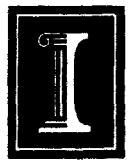


PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

LIBRARY TRENDS
SUMMER 1999
VOLUME 48, NUMBER 1, 1-281

Knowledge Discovery in Bibliographic Databases

Jian Qin
and
M. Jay Norton

Issue Editors

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

LIBRARY TRENDS

Library Trends, a quarterly thematic journal, focuses on current trends in all areas of library practice. Each issue addresses a single theme in depth, exploring topics of interest primarily to practicing librarians and information scientists and secondarily to educators and students.

Library Trends is published four times annually—in summer, fall, winter, and spring—by the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820-6211.

Subscriptions: Institutional rate is $85 per volume (plus $7 for overseas subscribers). Subscriptions for an individual are $60 (plus $7 for overseas subscribers). Registered students may subscribe for $25 (plus $7 for overseas subscribers). Individual issues are $18.50 (shipping included); back issues other than those from the present year are $10 (plus shipping). Claims for missing numbers should be made within six months following the date of publication. All foreign subscriptions and orders must be accompanied by payment.

Address orders to: University of Illinois Press, Journals Department, 1325 S. Oak Street, Champaign, IL 61820. For out-of-print issues, contact Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346. **Postmaster:** Send change of address to University of Illinois Press, 1325 S. Oak Street, Champaign, IL 61820-6903.

This journal is abstracted or indexed in *Library and Information Science Abstracts, Current Contents, Current Index to Journals in Education, Information Science Abstracts, Library Literature, PAIS,* and *Social Sciences Citation Index.*

Procedures for Proposing and Guest Editing an Issue of Library Trends

We encourage our readers to submit ideas for future *Library Trends* themes; issue topics are developed through recommendations from members of the Publications Committee and from reader suggestions. We also encourage readers to volunteer to be issue editors or to suggest others who may be willing to be issue editors.

The style and tone of the journal is formal rather than journalistic or popular. *Library Trends* reviews the literature, summarizes current practice and thinking, and evaluates new directions in library practice. Papers must represent original work. Extensive updates of previously published papers are acceptable, but revisions or adaptations of published work are not sought.

An issue editor proposes the theme and scope of a new issue, draws up a list of prospective authors and article topics, and provides short annotations of the article's scope or else gives a statement of philosophy guiding the issue's development. Please send your ideas or inquiries to F. W. Lancaster, Editor, Publications Office, 501 E. Daniel Street, Champaign, IL 61820-6211.

LIBRARY TRENDS

Summer 1999 48(1) 1-281

Knowledge Discovery in Bibliographic Databases

Jian Qin
and
M. Jay Norton

Issue Editors

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

This Page Intentionally Left Blank

Knowledge Discovery in Bibliographic Databases

CONTENTS

Introduction

JIAN QIN AND M. JAY NORTON

IN THE PAST FEW YEARS, A NUMBER OF research journals in library and information science have published review articles or special issues on knowledge discovery and data mining (Raghavan et al., 1998; Trybula, 1997; Vickery, 1997). These publications have primarily discussed background, scope and terminology, methods and techniques, and tools related to the topic from orientations other than library and information science. Research publications in library and information science have been implicitly related to knowledge discovery in databases (KDD) in terms of methods and techniques, though many of them did not use the terminology "knowledge discovery in databases" explicitly. This issue is devoted to aspects of KDD that are relevant or reflective of the field of library and information science.

Knowledge discovery in databases uses a variety of methods to evaluate data for relevant relationships that could yield new knowledge. According to Fayyad et al. (1996): "KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process" (p. 39). Data mining essentially focuses on identifying patterns previously not recognized and is considered only one component of the discovery process. KDD encompasses a growing collection of techniques, from a variety of disciplines, for investigating data to extract knowledge. The methods employ a broad combination and application of human expertise and information technology. "KDD comprises

Jian Qin, School of Information Studies, Syracuse University, 4-206 Center for Science and Technology, Syracuse, NY 13244
M. Jay Norton, School of Library and Information Science University of Southern Mississippi, Hattiesburg, MS 39406-5746
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 1-8

many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations" (Fayyad et al., 1996, p. 41). KDD investigates databases to identify patterns of association, clusters, and rules but it requires significant rigor—not all patterns are real or meaningful. The presence of patterns may be meaningless and statistically insignificant. The successful use of data mining in KDD involves "data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of results of mining" (Fayyad et al., 1996, p. 39).

On a fundamental level, library and information services have been involved in component processes similar to the current definition of KDD. Practitioners and researchers in library and information science have expended significant resources—intellectual and physical—on investigating and developing methods to identify and exploit patterns within information entities. These methods are used to generate classification schemes and organization systems for information retrieval and to address often poorly expressed information needs of users. In seeking ways to provide better access to information, the field has attempted to determine characteristics of relevance in query construction and investigated methods for improving document retrieval. KDD studies in library and information science that Fayyad (1996) identifies relate to and drive KDD and include statistics, areas of artificial intelligence, pattern recognition, visualization, intelligent agents for distributed and multimedia environments, machine learning, databases, management information systems, knowledge acquisition, information retrieval, and digital libraries (p. 23). In some fields, KDD is interpreted as applying whatever computer rigor and capability is available for extracting information from databases of all constructions, while in others it may have fewer technological implementations but the same desired outcome—i.e., the discovery of useful information (Fayyad, 1996). Practitioners of library and information science may see themselves more as intermediaries, or part of the process, though researchers in the field may see themselves as discoverers. KDD is, and will continue to be, a complex, multidisciplinary, interdisciplinary arena requiring both practitioners and researchers. As the field continues to develop, it will be interesting to compare the disjointed records of some of the disciplines to determine if the same issues arise—i.e., standardization of database construction, development of algorithm rules related to specific topic collections, and questions of subject expert classification versus external classification systems.

The thirteen articles included in this issue characterize a combination of the knowledge discovery in data process components; the emerging information technology; and the established information methods such as classification, citation analysis, and indexing and abstracting. Norton's article begins the issue by giving an overview of what KDD is and what

problems researchers face in KDD applications. She reviews the relationship between databases and knowledge discovery and the factors affecting the database quality that in turn impact the reliability and validity of KDD results. The article emphasizes that KDD is not at all a finished product, nor is it a panacea for all the research interests or ills of the database universe. In the face of many challenges in KDD, human involvement plays a vital role in the process.

Kwasnik discusses the relationship between knowledge representation (as manifested in classifications) and the processes of knowledge discovery and creation. While classifications categorize and interrelate domains and branches in the knowledge system, the classification process has potential to enable or constrain knowing something or discovering new knowledge about something. To demonstrate this, Kwasnik first describes the structures of a classification, including hierarchies, trees, paradigms, and faceted analysis with the goal of identifying how these structures serve as knowledge representations and in what ways they can be used for knowledge discovery and creation. When one considers that classification is built on known information, then KDD and classification takes on a new construction. Since a large part of KDD attempts to identify information that has previously been overlooked or unavailable, KDD will in itself affect classification. Basic constructs will remain the same but the underlying knowledge foundations that we apply to classification of an information entity will have to become more fluid in order to serve and be served by KDD. Kwasnik concludes that classification systems that are too rigid will not be applicable in the long term and may actually be detrimental to future knowledge discovery.

Swanson and Smalheiser report in their article the recent development in their text linkage discovery tool, Arrowsmith, a software program that draws upon expert knowledge in discovering implicit links among documents that have accumulated from the research done by Swanson (for a list of publications, see Swanson and Smalheiser's article in this issue of *Library Trends*) for more than a decade. Swanson's theory is based on the analogy that, if an article reports an association between substance *A* and some physiological parameter *B* while another reports a relationship between *B* and disease *C*, and a link between *A* and *C* via *B* has not been published previously, then to bring together the separate articles on *A-B* and *B-C* may suggest a novel *A-C* relationship of scientific interest. Arrowsmith is designed to develop systematic methods for discovering the undiscovered implicit relationships within the biomedical literature. It filters the text, matches phrases and concepts, and identifies potentially complementary items as pairs, which the researcher then analyzes for possible relationships. The software enables the investigator to evaluate relatively large bodies of data from a variety of aspects in a knowledge discovery mission. Recurrent in this work is the role of the human

investigator—Arrowsmith is a tool for discovery but is not the discoverer. Software such as Arrowsmith enlarges the scope of our view but does not replace the human analysis. The ability to scrutinize substantial databases to extract potentially revealing, and previously unnoted, information is a result of improving technology that portends tremendous benefits.

The discussion by Cory describes his experiment using Swanson's methodology to investigate, through document retrieval, whether three philosophers from different times in history were influenced by one another. Discovering the undiscovered text linkages among documents is less problematic in the biomedical literature than in the humanities because the technical terminology is usually explicit and precise while the humanities literature often abounds with synonyms. Will Swanson's methods be applicable to the humanities literature and yield the same type of links among the humanities documents? The literature inquiry about three philosophers from different historical times showed evidence of influences that the third philosopher received from the second and the second received from the first. The search was able to identify publications showing that the third philosopher was also influenced by the first. While Cory's method departed significantly from Swanson's work due to the idiosyncrasies of the humanities language, this experiment nevertheless linked logically-related citations that were bibliographically unlinked. His article also discusses the problems in discovering hidden knowledge in humanities databases because of the nature of humanities research and the language used in the humanities literature.

Small demonstrates in his article how citation links can be used to map scientific passages crossing disciplinary boundaries. Both Swanson and Cory's studies indicate the importance of analogy in discovering covert relationships among documents. While their methodology focuses on "recurring" terms or names that are shared by the documents found, Small maintains that citation links represent a more direct author-selected dependency than vocabulary sharing. This allows citation links to be used to establish frequency patterns of co-citation or bibliographic coupling, and thus they are more objective in studying the unity of science from a global perspective. In his study, Small generated a path by selecting economics as the starting field and astrophysics as the destination field. The citation links reveal that this path traverses the fields of economics, psychology, neuroscience, biomedicine, genetics, chemistry, earth science, geoscience, semiconductors, lasers, and physics. The co-citation passage from economics to astrophysics embraces interdisciplinary boundary spanning, such as psychiatry to neuroscience, neuroscience to immunology, and biology to biochemistry.

A similar analogy to Swanson's can be made in co-citation passage analysis that, if *A* is in the starting field, *C* in the destination field, and *B* the shared concept/method, then *A* is to *B* as *C* is to *B*. Small suggests

that, in future retrieval systems, a user could pick two topics or documents and generate a path of documents or topics that connect them, which could be used for information discovery and hypothesis generation.

The discussion by Qin addresses the problem of preprocessing and cleansing textual data for discovering semantic patterns in keyword frequency distributions. Keywords that are used as indexing terms in bibliographic records are semi-structured data. One challenge in mining such semi-structured data is to transform these into the types and structures suitable for statistical calculations and modeling. As semantic pattern analysis needs accurate data to draw valid and reliable conclusions, all the idiosyncrasies existing in natural language, including suffixes, different spellings for the same word, and synonyms, need to be normalized. Qin proposes the use of brief text codes to normalize the keywords while maintaining their original meaning. Besides the methodological aspect of mining bibliographic data, the frequency distribution patterns in the keyword data set suggest the existence of a common intellectual base with a wide range of specialties and marginal areas in the subject area studied. In normalizing the frequency of keyword occurrences, Qin found that the degree of keyword scattering at a certain region—i.e., keyword density—can be measured by the ratio of the number of unique keywords to the number of ranks at which the unique keywords occurred. The resulting values show a difference oftentimes between the specialty and marginal keyword regions. The semantic pattern analysis of the keywords from bibliographical coupling shows a possibility that simple semantic processing of natural language (keywords extracted from citation titles in this case) may be programmed into information retrieval tools for providing "analyzed" search results to users.

In his article, He reviews the development, applications, and advances made in co-word analysis during the last two decades. Though still developing as a technique, co-word analysis has been used in a variety of situations. Conjunct with its use is the recognition of one of its shortcomings—i.e., the assignment of keywords and indexing terms by indexers or database producers rather than the authors of the material. However, improving technology may allow the application of co-word analysis to full text to determine the appropriate keywords and indexing terms. It is through the application of such methods as co-word analysis that it is possible to identify problems in the construction of the databases and to consider the impact of indexers' choices on future retrieval and understanding of the semantic structures of a discipline. The creation of knowledge discovery methods also results in knowledge discovery as it highlights issues, concerns, and activities not previously scrutinized under other methods.

The articles mentioned above have concentrated on finding document content linkages and semantic patterns from the data available in

bibliographic databases. As digital documents grow exponentially, needs for organizing and retrieving these documents also arise. How can the subject content of digital full-text documents be represented effectively for retrieval purposes? What characteristics exist in these digital documents? How can these characteristics be organized and implemented in information systems to assist people in knowledge discovery? The following contributions address these questions from three different perspectives.

Ahonen's article analyzes digital document collections by identifying descriptive or meaningful word sequences that may be used in a variety of knowledge discovery missions. In extracting frequent word sequences from full-text documents, Ahonen posits that there may be common measures of relevance that can be detected by examining characteristics of word sequences. Her discussion provides a detailed account of the methods involved and demonstrates the potential of word sequence evaluation for knowledge discovery. Patterns in word sequences may be produced, based on a combination of pre- and post-processing linked to the specific application and frequency relations defined by rule sets and weight systems. The patterns may suggest areas of further investigation, be used to preevaluate a document's relevance without examining the whole document, or provide context for one not familiar with the document collection. The subject expert might also discern new information from the sequence associations or patterns.

Chowdhury presents a selection of cases where template mining has been successfully applied for information extraction from digital documents. Additionally, he reports on template use in Web search engines conducting information retrieval rather than information extraction. The initial distinction is that information retrieval attempts to locate relevant documents from collections while information extraction attempts to pull relevant information from documents. Though these are degrees of retrieval to some, the difference can be significant. The templates designed to assist the Web user in searching are created by expert searchers who organize information into groups and topics that are used to create the template structure for the less experienced user to plug into. The template is used to locate documents. The templates he ultimately focuses on have the potential advantages of authors using the template system to implement a more controlled method for creating document surrogates and digital document description to better enable information extraction from the documents, not just the collection. Not proposing a single all-purpose metadata format at this time, he suggests further research and investigation into what would be the most appropriate format.

Desai et al. developed a virtual library indexing and discovery system named CINDI (Concordia INdexing and DIscovery System) that allows authors of digital documents to describe their document via completion

of a semantic header and use of an expert registry subsystem. An appealing aspect of knowledge discovery in databases involves locating knowledge that might otherwise be overlooked. The Internet search engines often suffer from a lack of organization and consistency in the collection space. An extraordinary number of retrieved documents preclude appropriate evaluation and tend to result in missed opportunities rather than recovered data. Some of the more complex endeavors of KDD are seeking ways to access legacy data that are not organized consistently. Current developers of data warehouses are encouraging more standardization as future redress for the problem (Bontempo & Zagelow, 1998). The header contains the metadata used by the searching systems to determine the appropriateness of retrieving that resource. CINDI provides assistance comparable to the expert cataloger or indexer for the author, addressing the shortcomings of many current search engines via better metadata description. The outcome of the use of CINDI should be a significant improvement in the ability of searchers to locate materials relevant to their inquiry. This knowledge discovery approach begins with the initial document, which will produce improved results in the future. It will rely more on known relationships than unknown but should enhance retrieval of related documents.

Pinto and Lancaster offer a new view on abstracts and abstracting—i.e., that the quality of abstracts is extremely important in knowledge discovery tasks. Because of the dual roles of content descriptor and retrieval tool, abstracts must maintain the quality of accuracy, readability, cohesion/coherence, and brevity. However, the importance of these criteria is likely to vary depending on who will be reading the abstracts. For abstracts intended solely for search purposes, such criteria as readability and coherence/cohesion are not important, while other attributes are applicable in other ways. Pinto and Lancaster maintain that the increasing application of computers to text processing has not reduced the value of abstracts, and their value should not diminish as more critical or sophisticated operations, including those of knowledge discovery, are applied to the text.

In exploring knowledge from geospatial information systems (GIS), Yu demonstrates, through GeoMatch, a GIS-based prototype system for cartographic information retrieval, that coordinates data in MARC records can be processed to provide understandable and useful knowledge for users in selecting information relevant to their needs. GeoMatch is a graphic-based interface that mines the geographical data buried in MARC records and other geospatial sources and visualizes the new knowledge discovered in these data. Discovering knowledge in geospatial data is distinct from text information searching because it uses algorithms to convert the coordinates information into user-understandable and useful knowledge. The main contribution of GeoMatch is the quantitative analysis

of overlapping relationships in the retrieval process. Not only can it help users to more precisely define their information need and adjust the searching strategy, but also it can be used to rank the result. The KDD applications of this type have constructive implications for information retrieval.

Finishing out this issue is distinguished Professor Emeritus Herbert S. White, former dean of Indiana University School of Library and Information Science. In "Librarians and Information Technology: Which is the Tail and which is the Dog?" he discusses the role of library professionals in relation to the applications of database technology. He argues that some information technology has positioned the librarian contrary to the supportive service role that has surrounded the profession.

REFERENCES

Bontempo, C., & Zagelow, G. (1998). The IBM Data Warehouse Architecture. *Communications of the ACM, 41*(9), 38-48.

Borgman, C. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science, 37*(6), 387-400.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*(5), 20-25.

Fayyad, U. M., & Stolorz, P. (1997). Data mining and KDD: Promises and challenges. *Future Generation Computer Systems, 13*(2-3), 99-115.

Fayyad, U. M.; Piatetsky-Shapiro, G.; & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Frawley, W. J.; Piatetsky-Shapiro, G.; & Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In G. Piatetsky-Sharpiro & W. J. Frawley (Eds.), *Knowledge discovery in databases* (pp. 1-27). Cambridge, MA: AAAI Press.

Raghavan, V.V.; Deogun, J. S.; & Sever, H. (Eds.). (1998). Special topical issue: Knowledge discovery and data mining. *Journal of the American Society for Information Science, 49*(5).

Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science & Technology, 32*, 197-229.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Knowledge Discovery in Databases

M. JAY NORTON

ABSTRACT

KNOWLEDGE DISCOVERY IN DATABASES (KDD) revolves around the investigation and creation of knowledge, processes, algorithms, and the mechanisms for retrieving potential knowledge from data collections. Related issues include data collection, database design, the description of entries in the database using the most appropriate representation, and data quality. This article is an introductory overview of knowledge discovery in databases. The rationale and environment of its development and applications are discussed. Issues related to database design and collection are reviewed.

INTRODUCTION

Development of techniques to investigate databases, or the contents of databases, is of significant interest. As data storage space becomes less expensive, data collection as a tool has become more accessible and more used. Organizations are literally stockpiling data in warehouses for future investigation. Research is being done to ascertain if there are patterns, not just within databases but within documents and disciplines, that contribute to knowledge retrieval.

Every discipline has borders that expand and contract with the practical and intellectual adventurism of its members. As the collective knowledge base has grown, it is apparent that aspects of one field cross into many other fields. The evolution of information technology also provides a bridge across disciplines—in its theories and applications to various

M. Jay Norton, School of Library and Information Science, The University of Southern Mississippi, Box 5146, Hattiesburg, MS 39406-5146
LIBRARY TRENDS, Volume 48, Number 1, Summer 1999, pp. 9-21

disciplines. Knowledge discovery in databases (KDD) is another manifestation of the expansion of investigative tools across fields of interest and applications.

Many disciplines contribute to the undertaking of KDD. Some are more cognizant than others of the many factors involved with data collection. This article is an overview of knowledge discovery in databases. Discussion of recurring concerns from different perspectives about the collection, classification, and quality of data related to applications of KDD is presented.

Databases and Knowledge Discovery

Dramatic improvements in information technology have encouraged the massive collection and storage of data in all areas from commerce to research. From operational databases where personnel data are kept; to transactional systems that track sales, inventory and patron data; to full-text document databases and more; databases are growing in size, number, and application. The enormous increase in databases of all sizes and designs is evidence of our ability to collect data, but it also creates the necessity for better methods to access and analyze data. Human capacity to handle the data available in these databases is not adequate for timely examination and analysis. Technology presents opportunities to maximize the use of these data in an economical and timely fashion. Attempts to improve the search and discovery processes when dealing with databases have generated significant interest across many fields resulting in a multidisciplinary approach. Knowledge discovery in databases employs diverse fields of interest including statistics, computer science, and business, as well as an array of methodologies, many still evolving: machine learning, pattern recognition, artificial intelligence, knowledge acquisition for expert systems, and more. Knowledge discovery in databases revolves around the investigation and creation of knowledge, processes, algorithms, and mechanisms for addressing the retrieval of potential knowledge. An important component of this activity is identification of patterns or trends, from metadata through, and including, the semantic level, which suggest an entity's relationships. KDD techniques have been successful with large-scale scientific databases, notably in astronomy to classify sky objects. In addition, techniques have been used in medical, environmental, political, and census research. Other applications have been made with industrial and business-oriented databases in marketing, finance, manufacturing, and Internet agents (Fayyad et al., 1996, pp. 37-38; Vickery, 1997, pp. 107-08).

The phrase "knowledge discovery in databases" is attributed to a 1989 workshop on KDD (Fayyad, 1996). The phrase was intended to clarify that the end result of investigating data should be the discovery of usable knowledge and to differentiate KDD as a whole process, not just one of its

components—i.e., data mining (Fayyad et al., 1996, p. 39). Knowledge discovery in databases encompasses all the processes, both automated and nonautomated, that enhance or enable the exploration of databases, large and small, to extract potential knowledge. The most commonly referenced component of these processes has been data mining which involves activities oriented toward identifying patterns or models in data representation, classification, semantics, rules application, and so on (Fayyad et al., 1996).

Emphasis that KDD is a whole process is intended to clarify that knowledge seeking in data collections involves intellectual and technological undertakings designed to seek useful knowledge and not merely stir data. Certain basic premises underlie these efforts: (1) knowledge is a relevant term rooted in individual information bases and needs; (2) finding patterns in data is not equivalent to discovering information; (3) data mining, to be effective, must be structured; (4) results of any discovery activity have to be evaluated within a context; (5) search mechanisms of this type of inquiry may require substantial iteration; and (6) many aspects of KDD are dynamic and interactive in application. While some facets of KDD are best served by technology, the ultimate evaluators and discoverers are the human agents generating the initial queries and directing the process (Fayyad, 1996; Fayyad et al., 1996).

Legacy and Design

Though one mission of KDD is to automate as many of the basic processes as possible, several factors impede progress in this sector. Intelligent data analysis techniques are still not sophisticated enough to resolve some data problems without intervention. The methods for identifying information appropriate to include in a database, adding it to the classification and organizational scheme of the database, and providing access points for retrieval are neither trivial nor uniform. Design and implementation of databases has relied on the purpose, scope, data characteristics, and technical limitations of the organization sponsoring the enterprise. The vitality of these databases has been dependent on the imposition of appropriate criteria for inclusion, characterization, and maintenance. Legacy databases designed for specific organizational tasks are rarely uniform in structure within a given enterprise, nor is there consistent data quality, representation, or depth. This diminishes the possibility of generalizing even tasks which are common to each discovery effort as the description of the database has to be customized, and variations in the construction and quality of the data accommodated (Raghavan et al., 1998; Deogun & Sever, 1998; Fayyad, 1996).

Databases are organized collections of data. They can typically be separated into reference or source databases. According to Rowley (1992): "*Reference databases* refer or point the users to another source (such as a

document, an organization or an individual for additional information or the full text of the document" (p. 14). These databases may contain citations, abstracts, addresses, and directory type information that allows the user to locate other resources. "*Source databases* contain the original source data" (p. 15) and may include a combination of numeric and text data such as corporate reports, stock information, pure numeric data such as statistics, or full-text documents (Rowley, 1992). It is common to use surrogates, which allow for locating information about an entity without having to interact directly with the primary entity or full-text data as a method to identify and manipulate the data. Such surrogates could be a title, a citation, an abstract, or any attribute that may be identified and associated with a specific entity. The surrogates may be what is to be manipulated in order to understand or react to the entities. An inventory database could be the collection of information that reflects the holdings of the business, the movement of the inventory, the stock, or the vendors. Sorting these data can yield information about inventory levels or the speed in which a vendor responds to orders. The data may be a surrogate or a representation for activities. Properly configured, it may be possible to use the database to model activities. For example, if a vendor takes longer to fulfill an order than another, it might be advisable to have an earlier reorder date attached to the stock of that vendor. Full-text databases may also contain full and complete documents and may or may not have a metadata descriptor set that includes subject fields, though most will have minimal fields such as author, title, and publication data. Databases are collections based on some relationship—maybe as basic as membership in the collection—that causes them to be placed in common or related files. The attributes that describe the entity are portions of an overall structure that should optimize the collection of data relevant to, and descriptive of, the entities. Consider a checkbook page; there is a column for a date, check number, item description, transaction amount, and transaction. Each column is an attribute and is intended to contain information that describes the checking transaction. Each row forms a single record—i.e., the fields or attributes that describe one entity. The entity is the checking transaction associated with one check number. When attributes appropriately and adequately describe an entity, it provides a better understanding of the entity and may reveal information about one entity's relationship to another. Information limitations, as well as constraints of space and money, impact database design. Collection of data may be based on weighing cost and need versus alternative resources and attempting to serve the most critical information needs. Designing databases takes into account what information resources the organization might require as well as the costs involved in time and technology in acquiring the data. Costs may be related to whether it is incidental to other activities, such as purchasing history collected at the checkout counter as part

of the inventory control program, or full text of documents acquired as part of the publishing process versus directed collection such as surveying. Designing database systems usually involves modeling the information environment and information mission for which the database is being implemented. Information limitations and costs are related to what is known and not known about the users, the environment, and the corpus of resources they might require now and in the future. Information needs, environments, and cost factors change over time.

A consideration of KDD database design and cost is data quality. The accuracy of the data's representation of the entity and environment from which it originated, as well as its currency, are factors of data quality. Orr (1998) uses the theoretical framework of the feedback-control system (FCS) to define data quality as "the measure of the agreement between the data views presented by an information system and the same data in the real world" (p. 67). His position is that data entered into databases and left unused for periods of time, without feedback, may become stagnant in comparison to what occurs to the entity the data originally represented. The lack of, or the failure to apply, feedback to data creates a discontinuity between static data and the continually changing world. For example, if the age of a person is recorded in a database but not the birth date or an aging algorithm, the age data remain the same and quickly become inaccurate in reflecting the age characteristic of the entity. If the attribute of age is not used, the error may go unnoticed, another aspect of Orr's contention that data that are unused may lose representativeness. Not utilizing data may result in not recognizing it has been erroneous or that it may never have been useful initially. Further, if data are used but do not specify criteria surrounding its acquisition, use, and maintenance, the value of the data will be decreased. Failing to include attributes to handle name changes, or failing to update a record when a name change is reported, reduces the accuracy of the record, may impede the location of the remainder of the record at a future time, and may miss data related to the name change pertinent to the record. Lack of rules governing the maintenance of data elements may cause reassignment of a field application without any overt documentation or clear history. For example, when it was discovered that there was no field to capture name changes, another field that did not seem much used might be informally redesignated as the name change field. At a later date, possibly using KDD techniques, it becomes apparent that the field in question was being used for recording the names of beneficiaries for specific insurance plans that were not grandfathered into the new plan. Data quality problems of this type will tend to multiply over time until the entire database's quality and usefulness is questionable.

What can happen to data in these situations can also happen to metadata. This perspective of data quality may be an issue of significant

importance in light of the trend toward data warehouses—i.e., if data are collected but unused, how accurate will it be by the time it is used (Orr, 1998)? A converse concern might be whether metadata records for documentary entities that contain subject descriptions should have the subject words modified to reflect new nomenclature or preferred subject terms? That is, what should happen if the environment of the information changes but the entity must remain the same? Should metadata change?

Intentional collection of data prior to identification of any specific purpose for it results from the recognition that information needs change over time and the data may be an unrecognized reservoir of knowledge. The emergence of data warehouses as a means to capitalize on an organization's data collection activities has potential as an advantage for KDD activities. "Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business" (Gardner, 1998, p. 54). If data warehousing is undertaken in a planned and logical manner, according to Fayyad et al. (1996), it could improve KDD opportunities and applications. With future KDD in mind, initial determination of how a data warehouse is designed, what attributes will be included, how the structures will be related or not will require more attention. Something to be considered is what information will be contained in the warehouse and how it can best be represented to require the least manipulation to access. If the warehousing organization will invest in uniform representation—methods for covering missing data and correcting errors—it will significantly decrease the preparation of data for KDD. Whether discussing databases or data warehouses, the underlying requirements to improve access are planning the collection, organizing, and rationally characterizing the structure for the best handling of the data with as much flexibility as possible. Some anticipation of what information will serve users in the future and how to provide access to the data without knowing what might be relevant in the future is the challenge (Sen & Jacob, 1998).

Currently, databases having KDD techniques applied to them were not necessarily built with this exploratory methodology in mind. Indeed, some of the techniques developed for KDD are in response to the lack of uniformity in database construction and omissions in retrieval capacity. Selection for inclusion in a database is based on user needs as they are identified in the construction process. If the construction process has not taken into account potential changes in information requirements, managerial decision path modifications, or new product data considerations, the database may have limited future value. The combination of currently serving users and forecasting what future services will be needed is a significant collection and design problem. Experience with legacy database data quality emphasizes the necessity of reviewing and improving planning and construction of databases and warehouses.

Classification

Historically, methods to provide access to the collected corpus of information have resulted in the imposition of artificial or controlled classification structures and languages. An example would be the development of classification systems such as Dewey Decimal or Library of Congress; both attempt to organize knowledge. Attached to these are subject heading or descriptor manuals—e.g., Library of Congress Subject Headings, MeSH (Medical Subject Headings), or specific thesauri. These tools suggest the classification and position in the hierarchy of knowledge of materials, permitting both assignment of subject and retrieval of subject by conformity to structured headings. Use of a controlled language in describing entities entered into a collection provides parameters to be employed in both building and searching. The subject headings and controlled language attempt to address the multiple layers of meaning which are part of language. The labeling of entities and meaning of words may change dramatically from discipline to discipline but also within subsets of disciplines and even over time. The imposed structure allows for information retrieval in relative proportion to the searchers' ability to manipulate the system and how well the information entered fits the structure. Use of controlled languages has resulted in using intermediaries to decode the systems imposed as the searchers were rarely those who classified the objects. The controlled structure also lent itself to application of information technology as uniform constructs are more easily manipulated by machines than natural language. Much machine searching currently relies upon matching input to some aspect of the database record. This may be simple and effective if the correct terms are entered into the searching algorithm—very resource intensive if the terms do not match and no internal algorithms allow for variations in the matching. When controlled language and related tools, such as indexes and thesauri, are implemented, it is possible to maximize the effectiveness of searches by using the controlled language. This naturally assumes that the language has been appropriately applied. When databases do not have controlled languages, the resources to search are more intensively expended, with varying results, dependent upon the searcher's ability to identify what terminology has been used to describe what they seek. The combination of machine limitations and the advantages of classification schemes impacted the design of early databases from both input and retrieval perspectives. Another method for organizing data for databases is embedded in the architecture of the database. By using a consistent structure exploiting the common attributes of the entities that are being entered into the database, it is possible to use the attribute structure for searching. For example, if the entity is an employee, then using attributes such as ID number, name, department of employment, supervisor, pension plan, or pay scale, could, if the searchware is properly designed, permit the searcher

to retrieve all employees from a given department and examine only their records or only the records of those on a specific pension plan. Early space and memory shortcomings restricted the amount and manner in which data could be stored. Data were "abbreviated" and arranged to maximize space savings. This resulted in using codes in the attribute fields and should have involved the application of value range rules. For example, the pension plan mentioned above may have been noted by a numeric code tied to a specific pension plan. In this way only a few bytes of space would be required to retain the information as long as somewhere there was a list (paper or electronic) of the code number associated with the pension. In a database of customer data, it might involve recording the inventory number of materials purchased rather than a textual description, or using a zip code as a region identifier rather than a street address. Applying rules to the content of the fields (attributes) would include specifying whether the whole name was in a single field or in two fields, and if there was a field size limit such that long names would be truncated and, if so, how; what date representation will be used—year first or last—and how many digits for the year? These concerns, coupled with the characteristics of the schema used, perpetuated the need for intermediaries. Now KDD is part of the intermediary force that can maximize the usefulness of such databases.

Classification and organization schemes are critical to any retrieval activity. To date these have been limited by technology, economics, knowledge, and tradition to selected access points usually identified by people who are not experts in the given discipline. Developing classification schemes to accommodate all knowledge has proven to be an evolutionary process. As understanding is gained as to the interrelatedness of our world, restrictive class structures have to be modified. Classification occurs at the database design level. Determining what attributes will describe an entity, the governing criteria, and the detail of description will affect what retrieval is possible. Seeking additional patterns that may be hidden within databases to generate new classification criteria via KDD is complex but less so than attempting to expand attribute descriptions to be complete classification structures, especially when some characteristics are not apparent without KDD. The ideal KDD evaluates data for trends or patterns that might be otherwise overlooked and, if statistical relevance is found, these may indicate subclasses or relationships. Such relationships might be used to further clarify and expand a database's value. Recognizing that this trend exists when it was previously unknown can provide new information value that poses new questions requiring further examination. Knowledge arises from prior knowledge.

Full-text searching algorithms for documents and textual databases are options but are still technologically cumbersome when working with any sizable database. Even when full-text searching becomes more reli-

able and economical, representations of documents will continue to be employed. Despite the power of web crawlers in recording and tracking pages, there is significant discourse about the use of metadata elements to represent pages. The bandwidth and time economy of identifying potentially desirable documents via surrogates, such as metadata or bibliographic record notations, is a significant savings. Full-text searching can be resource intensive and currently not particularly more effective than surrogate searches. Despite the advances in information technology, there are still difficulties with searching efficiently in large-scale databases. Seeking patterns in data is compromised if only portions of the data can be evaluated at a time, something that many searching algorithms overlook relative to large-scale databases. If the assumption is that all necessary data to consider is in memory, when it cannot physically be so, means detected trends or patterns are false results. Continued research and development to provide better algorithms for manipulating what may be tetrabytes of data in some rational manner is proceeding (Fayyad et al., 1996). Devising more complex sampling or modeling approaches using KDD techniques may yield some advantage. Certainly as the technology continues to advance, the application of pure brute processing power may be the answer.

The Processes

The actual processes are much more involved and complex than presented here. The following is a limited overview of what is a formidable undertaking. The application of KDD techniques requires substantial researcher involvement in determining the problem to explore and framing it within a meaningful context. Further, the investigator will, by necessity, be engaged in repetitive examination of the processes and results to direct or redirect the exploration. The discovery processes may dramatically influence the paths that research must follow.

Similar to any research endeavor, KDD requires defining the problem domain and acquiring underlying information relevant to the inquiry to identify the research path to follow. Establishing the parameters of the problem and determining the potential goal of the research is followed by the selection and possible extraction of the data set or subset to explore. Depending on the problem, a test set may be necessary to identify the best methodology. In fact, ascertaining the appropriate data to examine may in itself be a series of tasks and tests. The data set must then be prepared, and rules for dealing with missing data, erroneous data, redundant attributes, data corruption, and such must be determined and implemented. The problem of legacy databases—diverse platforms, errors introduced over time, changes in data entry procedures, poorly organized data, or technological limitations—must be adjusted for to enable statistical manipulation. (Perhaps the amount of preparation of data involved

in the data mining phase of knowledge discovery is why there has been so much focus on the data mining process.) The data set is then reduced or transformed, if appropriate, and/or standardized in representation and structure to enable manipulation, analysis, or modeling. There are a large number of possible algorithms to apply depending on the problem or the theorized pattern and the goal of the exploration. If patterns emerge, they must be analyzed, evaluated, and retested. Any or all of these processes may require repetition or modification along the way in response to difficulties encountered and findings that might influence the original theory. Selection of the methods to apply in a given situation is related to the intention of the research, amount and construct of the available data, as well as the quality of the data. There are no hard and fast rules governing the application of techniques beyond the appropriateness of one method to a particular domain or problem. Like any research effort, posing the correct or best inquiry will provide the best results (Fayyad et al., 1996). KDD is usually invoked to verify or discover; just pulling out data patterns is not sufficient. Some additional measures or tests are necessary to determine if data patterns have any value to the investigator, whether the pattern is an experimental phenomena or an actual recurring pattern. The expertise at this level of decision, whether there is any valuable meaning to derive from any detected patterns, has to come from the human investigator. By its nature, research requires human investment; curiosity is the fountain of knowledge discovery.

Reality Check

It would be a disservice to overlook two key points about knowledge discovery in databases. First, in the context presented, this is an emerging field, an evolving study, and not a finished product. Second, it is not a panacea for all the research interests or ills of the database universe. It does present momentous potential in its future incarnations in conjunction with evolving information technology, especially artificial intelligence areas. KDD is already demonstrating its value in its current state. Ongoing efforts to address the shortcomings of the data it examines and the technology employed will result in rapid advances. Many of the challenges facing KDD are the same as those which confront the entire information community:

- The multiple layers of data quality problems in databases resulting from design and implementation shortcomings present serious difficulty. When both modeling the information environment and the information mission have failed or have been incomplete, the database structure is inadequate to properly represent the data or to ensure consistency—e.g., a database containing student records that has no data field for name changes. In such case, a name change may not

be recorded or the new information lost, and any requests for the student with the new name will fail. When there are no rules for selecting data, no attempt to provide meaningful classification for the attributes, they do not describe the entities. For example, if customer names are to be kept in the database, but there is no rule for inclusion—such that sometimes names are entered as one field, sometime as two—it is possible for more than one record per customer to be created and for the search engine to be unable to match any of the records. If an inventory database has no timing attributes—that is, no date of inventory receipt or inventory decreases—then it is no more than a list of inventory which may or may not be present. Problems of data currency and accuracy in legacy databases, where the original collection of data may have been some time in the distant past and there have been no updates to the data, can damage the accuracy of the data. Lack of uniformity in the collection and loss of consistency as different entry systems changed field context can make a database unusable.

- Large databases with many attribute fields and variables pose complex and, as yet, unresolved computing and search difficulties. Memory management of these huge databases makes it difficult to analyze whole data sets at one pass, requiring different algorithms to perform analysis over smaller sets and still produce valid and reliable results. The sheer number of fields (attributes) in some databases make analysis extremely complex. Determining the influencing factors and fields to evaluate becomes a more sophisticated statistical problem as well as an information management problem.
- Increased complexity of relationships within databases requires more sophisticated search algorithms and more rigorous inspection techniques. This is a problem related to the depth of fields but also to the types of data collected. How datum fits into the database and its role in the information environment impacts our ability to analyze it. Why is datum included, and why is another not?
- Insufficient tools to incorporate prior knowledge into systems in more meaningful ways present special problems. What is the best way to make the available domain knowledge accessible to the search systems? Training systems to be expert systems is one approach, but this is still an evolving field and dependent upon human expertise. Further, as more is known, how should the system be adapted? Can we develop algorithms with sufficient robustness to adapt?
- Lack of historic platform integration and proprietary software restrictions contribute to the confusion and frustration of dealing with legacy databases. When databases are restricted to a specific platform and a specific software interface or program manipulation, it requires investment to adapt or overcome the barriers. Sometimes it means

re-entering the data in a less restrictive system, which increases the likelihood of data corruption.

- Certainly not the last nor the least challenge to KDD is the lack of information and knowledge about the human factors and roles in the construction, design, collection, classification, and retrieval related to databases (Fayyad et al., 1996). The construction of databases is a series of complex problems that include modeling the information environment as it is perceived and forecast at the time; identifying the appropriate information to incorporate into a database; and selecting the most representative attributes and the value limitations, to mention but a few. Each of these decisions is limited by technology—both the availability and the designers' ability to use it. Characterizing environments is a human activity that is embedded in belief systems, political perspectives, social concerns, and business acumen. The designers' background, comprehension of the information problem, the needs of the users, and the demands of the future all will impact the outcome of the database. How the construction is undertaken, and how the rules for validation and inclusion are composed and conveyed to the data collectors all affect the database contents. If the database design is exceptional and the interface used to enter data is poor, the result will be a poor database. The presence of thorough planning, construction, and implementation documentation is critical for those attempting KDD. Though often not available, such documentation could provide context to significantly improve the investigation. Even though computers collect data, the human actors design the database, create intellectual models for its construction and implementation and, ultimately, for its reinterpretation. Leaders in the KDD effort stress the role of human involvement in the retrieval processes. It is also a critical and little understood factor in the creation processes. Understanding the human cognitive processes involved in creating a search or determining what is the solution to a search is critical to the future of KDD. Better understanding of human cognition and pattern recognition could yield important clues to improved algorithms for computer cognition.

Conclusion

Emerging fields, new approaches, and knowledge discovery all herald change. Indeed, KDD does remold some aspects of research by implementation of a wide variety of tools from an array of disciplines; it advances the interrelatedness of the effort. The techniques have broad potential for application. Some aspects of KDD are indeed rediscovered knowledge. Some would argue that much of modern bibliometrics is kindred to KDD. Others might note that much of it is the application of basic statistics to another set of problems. It is clear that the techniques

are still very preliminary in their current applications, though many of the techniques have existed for some time. Legacy database design problems are bound to KDD because the techniques can detect some of them and because some of them complicate KDD. There is much more work to be done in this area. Clearly, more emphasis and research into designing databases and data warehouses is needed.

REFERENCES

Fayyad, U. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*(5), 220-225.

Fayyad, U.; Piatetsky-Shapiro, G.; & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Ai Magazine, 17*(3), 37-54.

Gardner, S. R. (1998). Building the data warehouse. *Communications of the ACM, 41*(9), 52-60.

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM, 41*(2), 66-71.

Pao, M. L. (1989). *Concepts of information retrieval.* Englewood, CO: Libraries Unlimited.

Raghavan, V. V.; Deogun, J. S.; & Sever, H. (Eds.). (1998). Introduction (In Special Topic Issues: Knowledge Discovery and Data Mining). *Journal of the American Society for Information Science, 49*(5), 397-402.

Rowley, J. E. (1992). *Organizing knowledge: An introduction to information retrieval.* Brookfield, VT: Ashgate.

Sen, A., & Jacob, V. S. (1998). Industrial-strength data warehousing. *Communications of the ACM, 41*(9), 29-31.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

The Role of Classification in Knowledge Representation and Discovery[1]

BARBARA H. KWASNIK

ABSTRACT

THE LINK BETWEEN CLASSIFICATION AND KNOWLEDGE is explored. Classification schemes have properties that enable the representation of entities and relationships in structures that reflect knowledge of the domain being classified. The strengths and limitations of four classificatory approaches are described in terms of their ability to reflect, discover, and create new knowledge. These approaches are hierarchies, trees, paradigms, and faceted analysis. Examples are provided of the way in which knowledge and the classification process affect each other.

INTRODUCTION

Developments in our ability to store and retrieve large amounts of information have stimulated an interest in new ways to exploit this information for advancing human knowledge. This article describes the relationship between knowledge representation (as manifested in classifications) and the processes of knowledge discovery and creation. How does the classification process enable or constrain knowing something or discovering new knowledge about something? In what ways might we develop classifications that will enhance our ability to discover meaningful information in our data stores?

The first part of the article describes several representative classificatory structures—hierarchies, trees, paradigms, and faceted analysis—with the aim of identifying how these structures serve as knowledge represen-

Barbara H. Kwasnik, School of Information Studies, 4-206 Center for Science and Technology, Syracuse University, Syracuse, NY 13244
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 22-47

tations and in what ways they can be used for knowledge discovery and creation. The second part of the discussion includes examples from existing classification schemes and discusses how the schemes reflect or fail to reflect knowledge.

Knowledge, Theory, and Classification

Scholars in many fields, from philosophy to cybernetics, have long discussed the concept of knowledge and the problems of representing knowledge in information systems. The distinction is drawn between merely observing, perceiving, or even describing things and truly *knowing* them. To know implies a process of integration of facts about objects and the context in which the objects and processes exist. Even in colloquial usage, knowledge about someone or something is always expressed in terms of deep relationships and meanings as well as its place in time and space. To *know* cars means not only understanding car mechanics but also knowledge of the interplay of the mechanical processes and perhaps even factors such as aesthetics, economics, and psychology.

The process of knowledge discovery and creation in science has traditionally followed the path of systematic exploration, observation, description, analysis, and synthesis and testing of phenomena and facts, all conducted within the communication framework of a particular research community with its accepted methodology and set of techniques. We know the process is not entirely rational but often is sparked and then fueled by insight, hunches, and leaps of faith (Bronowski, 1978). Moreover, research is *always* conducted within a particular political and cultural reality (Olson, 1998). Each researcher and, on a larger scale, each research community at various points must gather up the disparate pieces and in some way communicate what is known, expressing it in such a way as to be useful for further discovery and understanding. A variety of formats exist for the expression of knowledge—e.g., theories, models, formulas, descriptive reportage of many sorts, and polemical essays.

Of these formats, science particularly values theories and models because they are a "symbolic dimension of experience as opposed to the apprehension of brute fact" (Kaplan, 1963, p. 294) and can therefore be symbolically extended to cover new experiences. A theory thus explains a *particular* fact by abstracting the relationship of that fact to other facts. Grand, or covering, theories explain facts in an especially eloquent way and in a very wide (some would say, universal) set of situations. Thus, Darwinian, Marxist, or Freudian theories, for example, attempt to explain processes and behaviors in many contexts, but they do so at a high level of abstraction. There are relatively few grand theories, however, and we rely on the explanatory and descriptive usefulness of more "local" theories—theories that explain a more limited domain but with greater specificity.

Classification as Knowledge Representation

How are theories built? How does knowledge accumulate and then get shaped into a powerful representation? There are, of course, many processes involved, but often one of them is the process of classification. Classification is the meaningful clustering of experience. The process of classification can be used in a formative way and is thus useful during the preliminary stages of inquiry as a heuristic tool in discovery, analysis, and theorizing (Davies, 1989). Once concepts gel and the relationships among concepts become understood, a classification can be used as a rich representation of what is known and is thus useful in communication and in generating a fresh cycle of exploration, comparison, and theorizing. Kaplan (1963) states that "theory is not the aggregate of the new laws but their connectedness, as a bridge consists of girders only in that the girders are joined together in a particular way" (p. 297). A good classification functions in much the same way that a theory does, connecting concepts in a useful structure. If successful, it is, like a theory, descriptive, explanatory, heuristic, fruitful, and perhaps also elegant, parsimonious, and robust (Kwasnik, 1992b).

There are many approaches to the process of classification and to the construction of the foundation of classification schemes. Each kind of classification process has different goals, and each type of classification scheme has different structural properties as well as different strengths and weaknesses in terms of knowledge representation and knowledge discovery. The following is a representative sample of some common approaches and structures.

Hierarchies

We have inherited our understanding of hierarchical classifications from Aristotle (Ackrill, 1963), who posited that all nature comprised a unified whole. The whole could be subdivided, like a chicken leg at the joint, into "natural" classes, and each class further into subclasses, and so on—this process following an orderly and systematic set of rules of association and distinction. How do we know what a natural dividing place is, and how do we arrive at the rules for division and subdivision? According to Aristotle, only exhaustive observation can reveal each entity's true (essential) attributes, and only philosophy can guide us in determining the necessary and sufficient attributes for membership in any given class. In fact, according to Aristotle's philosophy, it is only when an entity is properly classed, and its essential properties identified, that we can say we truly *know* it. This is the aim of science, he claims—i.e., to unambiguously classify all phenomena by their essential (true) qualities.

While Aristotle's legacy is alive in spirit in modern applications of classification, most practitioners recognize that a pure and complete hierarchy is essentially possible only in the ideal. Nevertheless, in knowledge

domains that have theoretical foundations (such as germ theory in medicine and the theory of evolution in biology), hierarchies are the preferred structures for knowledge representation (see, for example, the excerpt from the Medical Subject Headings [MeSH] in Figure 1).

EYE DISEASES
- CONJUNCTIVAL DISEASES
  - CONJUNCTIVAL NEOPLASM
  - CONJUNCTIVITIS
    - CONJUNCTIVITIS, ALLERGIC
    - CONJUNCTIVITIS, BACTERIAL
      - OPHTHALMIA NEONATORUM
      - TRACHOMA
    - CONJUNCTIVITIS, VIRAL
    - KERATOCONJUNCTIVITIS
    - REITER'S DISEASE
- CORNEAL DISEASES
  - ETC.

Figure 1. Hierarchy: Excerpt from MeSH (Medical Subject Headings).[2]

Based on the MeSH excerpt in Figure 1, note that hierarchies have strict structural requirements:

- ***Inclusiveness.*** The top class (in this case, EYE DISEASES) is the most inclusive class and describes the domain of the classification. The top class includes all its subclasses and sub-subclasses. Put another way, all the classes in the example are included in the top class: EYE DISEASES.
- ***Species/differentia.*** A true hierarchy has only one type of relationship between its super- and subclasses and this is the generic relationship, also known as *species/differentia*, or more colloquially as the *is-a* relationship. In a generic relationship, ALLERGIC CONJUNCTIVITIS *is a* kind of CONJUNCTIVITIS, which in turn *is a* kind of CONJUNCTIVAL DISEASE, which in turn *is a* kind of EYE DISEASE.
- ***Inheritance.*** This requirement of strict class inclusion ensures that everything that is true for entities in any given class is also true for entities in its subclasses and sub-subclasses. Thus whatever is true of EYE DISEASES (as a whole) is also true of CONJUNCTIVAL DISEASES. Whatever is true of CONJUNCTIVAL DISEASES (as a whole) is also true of CONJUNCTIVITIS, and so on. This property is called *inheritance*, that is, attributes are inherited by a subclass from its superclass.
- ***Transitivity.*** Since attributes are inherited, all sub-subclasses are members of not only their immediate superclass but of every superclass above that one. Thus if BACTERIAL CONJUNCTIVITIS is a kind of CONJUNCTIVITIS, and CONJUNCTIVITIS is a kind of CONJUNCTIVAL DISEASE,

then, by the rules of transitivity, BACTERIAL CONJUNCTIVITIS is also a kind of CONJUNCTIVAL DISEASE, and so on. This property is called *transitivity.*

- ***Systematic and predictable rules for association and distinction.*** The rules for grouping entities in a class (i.e., creating a species) are determined beforehand, as are the rules for creating distinct subclasses (differentia). Thus all entities in a given class are like each other in some *predictable* (and predetermined) way, and these entities differ from entities in sibling classes in some predictable (and predetermined) way. In the example above, CONJUNCTIVAL DISEASES and CORNEAL DISEASES are alike in that they are both kinds of EYE DISEASES. They are differentiated from each other along some predictable and systematic criterion of distinction (in this case "part of the eye affected").
- ***Mutual exclusivity.*** A given entity can belong to only one class. This property is called *mutual exclusivity.*
- ***Necessary and sufficient criteria.*** In a pure hierarchical classification, membership in a given class is determined by rules of inclusion known as *necessary* and *sufficient* criteria. To belong to the class, an entity *must* have the prescribed (necessary) attributes; if it has the necessary attributes, this then constitutes sufficient warrant, and the entity must belong to the class.

Because of these formal properties, hierarchical classification schemes continue to have great appeal in knowledge representation and discovery for several reasons:

- ***Complete and comprehensive information.*** A hierarchical classification is usually a fairly comprehensive classification since all rules for aggregation and distinction must be made a priori. This means that, before the structure is established, the designer must know a great deal about the extent of the entities, their attributes, and the important criteria along which they are similar and different.
- ***Inheritance and economy of notation.*** The formalism of a hierarchy allows an economical representation of many complex attributes. Each attribute does not have to be repeated at each level but rather is inherited as part of the scheme. Much information can be "carried" by the hierarchical structure.
- ***Inference.*** For this reason, a hierarchy allows reasoning from incomplete evidence. If it can be established, for instance, that a patient has the symptoms of conjunctivitis (as defined by the necessary and sufficient criteria by which a set of symptoms is given this label), then it is possible to know also that, as a kind of eye disease, conjunctivitis will share properties with other eye diseases. This is especially useful if the shared criteria are not obvious or easily observable. For example, if, by observation and comparison with other animals, you assess that

an animal is a kind of *cat*, which is a kind of *mammal*, you can infer and predict that, if it is a female, it will reproduce by bearing live young and breast feeding its babies, even though these cat-like characteristics may not be immediately evident.

- ***Real definitions.*** Hierarchical classification enables *real definitions*, which are considered by many to be superior to other types of definitions because they provide a way of expressing how an entity is *like* something, and also how it is *different* in some important way. For instance, consider the definition: "A bachelor is an unmarried man." A bachelor is a man; therefore he shares all the characteristics of men. Men can be married or unmarried. A bachelor is of the "unmarried" type of man. The strength of this definition, as a definition, lies in its ability to succinctly describe a complex of attributes of affinity and an important aspect of distinction. Two alternative definitional strategies to real definitions are to list attributes one by one or to point to exemplars ("See that guy? He's a bachelor." "See that other guy? He's a bachelor too"). A real definition is often the more efficient way of describing the nature of the entity and the boundaries of where, by definition, that entity ends.
- ***High-level view and holistic perspective.*** If the criteria by which the classificatory structure is built are theoretical in the sense that they reveal fundamental and meaningful distinctions, then the classification scheme as a whole provides a visualization of the phenomena it is representing. Such a birdseye perspective enables recognition of overall patterns and anomalies, interesting or problematic relationships, and so on. A holistic high-level view is often a trigger for knowledge generation, allowing the researcher to step away from the individual instances to see them as they fit into a larger context.

Not every knowledge domain lends itself to being represented by a hierarchy, however. While hierarchies are desirable for their economy of notation, the richness of description, and the incorporation of knowledge about relationships, they are also problematic for a number of reasons:

- ***Multiple hierarchies.*** At the top of the list is the fact that, from our modern (non-Aristotelian) perspective, we no longer view the world as having only *one* reality—i.e., one way of being parsed neatly at the joints. Most phenomena are understood to have several, perhaps overlapping, but separate sets of attributes and relationships, depending on the context and goal of the representation. For instance, dogs are mammals and knowing they are mammals helps us understand their physiological selves. But dogs are also pets and as such belong to the domain of domesticated animals and human companions. Knowing this aids in understanding the social aspects of dog behavior in a particular context, as well as the larger social phenomenon of pet

ownership. This suggests that we must have separate classifications for "dogs as animals" and "dogs as pets" with perhaps some cross-links to show the connections in a tangled, or multihierarchical, structure. In any event, no one classification is able to capture all aspects of a particular domain.

- ***Multiple and diverse criteria.*** There seem to be some practical limits to how much information a hierarchy can bear in its structure before it becomes too complex. Consider the placement of *lions* in a classification of animals. Traditional zoological taxonomy, based on morphological attributes, places *lions* in with other *felines.* But consider the distinction between lions in the wild versus lions in zoos. Are they the same entity? A hierarchy is not well designed to accommodate distinctions made along two very different sets of criteria. While it is possible in theory to further subdivide each animal in a taxonomy of animals by whether it is in the wild or in captivity, such a representation becomes very cumbersome and repetitive. If a hierarchy is weighted down by too many perspectives and disparate rules for grouping and differentiation, it loses some of its power as a clear representation. One of the difficulties with traditional taxonomies of the living world, in fact, is its inability to accommodate the notion of "habitat." The representation of knowledge about living entities in ecological systems and over time is difficult in a hierarchy that requires conformity to the principle of mutual exclusivity. For example, in classifying *dinosaurs,* one must decide whether it is more useful to cluster a particular dinosaur under the domain of *prehistoric creatures* (thus using "age" as the defining factor) or to separate dinosaurs and classify each particular kind under the domain of *mammals, birds, reptiles,* and so on (thus focusing on their attributes as specific types of animals rather than on when they lived). To do both simultaneously is representationally difficult.
- ***Lack of complete and comprehensive knowledge.*** Since hierarchies attempt to be comprehensive and to show the relationship of all entities to each other in an overall structure, they require relatively complete knowledge of the domain in advance. In emerging fields, where the extent of the domain is not yet charted, where the relationships are not yet fully understood or defined, or where there is no theoretical framework on which to build the structure, a hierarchy is both difficult and inappropriate to build. It is not just a question of comprehensiveness. If a knowledge domain rushes into a hierarchical representation without adequate grounding or warrant, the result can be a representation that is misleading or skewed. Such representations can also lead to premature closure in terms of knowledge creation because a hierarchy implies clear boundaries and a complete set of criteria, while this may not in fact be the case. The sure sign of a

"premature" hierarchical structure is the need for a "miscellaneous" or "other" category into which the classifier places all those entities that do not fit into the logic of the classification system as specified.

- ***Differences of scale.*** In order to maintain the principles of transitivity and inheritance, all entities in a hierarchy must be at the same conceptual level of granularity. For example, in classifying the entity *beach*, it is possible to look at a beach from the global perspective and see "*an area of demarcation between land and sea*," or from the perspective of a human walking on it as "*sand, shells, seaweed, etc.*" or through a microscope as "*crystalline structures*"—same beach, different level of definition. Such differences in scale are not easy to accommodate in one classification. If combined into one structure, and especially if combined haphazardly, they weaken the integrity of the knowledge representation. This is because it is not clear at any given point in the classification which criteria of association and distinction are being invoked: beach as *land-mass type*, beach as *habitat*, or beach as *physical material*.
- ***Lack of transitivity.*** A hierarchy requires that attributes are passed on down the structure intact. So, if *A* is a subclass of *B*, and *B* is a subclass of *C*, then *A* is also a subclass of *C*. This neatness does not always translate into the way we humans perceive the phenomena around us. For example, we might all agree that *chairs* are a kind of *furniture*. Further, we might agree that *rocking chairs*, and *easy chairs*, possibly *stools*, and perhaps even *tree stumps* are a kind of *chair*—depending on the context. But, while most people would agree that a *stool* is also, therefore, a kind of *furniture* (thus conforming to the principle of transitivity), not all people would extend the inheritance and agree that a *tree stump* is a kind of *furniture*. In other words, somewhere in the chain of representation the rules change and not all the attributes of *furniture* get invoked in determining the nature of a *tree stump*. This situation leads to a knowledge representation that subtly shifts. As a consequence, it is not possible to use such a representation as a reliable source of inference.
- ***Rules for class inclusion are too strict.*** Entities do not always conform to the *necessary-and-sufficient* criterion. In a pure hierarchy, entities must belong unambiguously to a class. If they possess all the necessary attributes, they are in; if they lack any of the attributes, they are precluded from membership. In a hierarchy, each member of a class is therefore as good a representative of its class as any other. Unfortunately, human beings do not perceive things quite so neatly. Entities can belong to a class more or less. The criteria for inclusion might fit one entity better than they do another. One entity might be a better representative of a class than another. For instance, most people think a *robin* is closer to the prototype of a *bird* than is a *penguin*. Put another

way, *penguins* are not unambiguously a member of the *bird* class, even though they may in fact possess all the necessary and sufficient attributes of *bird*-ness but not to the same perceived degree as a *robin* does. Furthermore, entities in a class may share some attributes in common with each other, but not all might share the *same* attributes. Thus, in my family, there may be a distinctive nose, distinctive eyebrows, and a distinctive smile, but not all members must have *all* these attributes to be perceived as showing a family resemblance. Finally, an entity may belong to one class under one set of circumstances, and to another class under another set of circumstances, or to both simultaneously. One can be both a *parent* and a *student* or sometimes a *parent* and sometimes a *student*. It is possible to be sometimes a better exemplar of a *student* (closer to the prototype), while at other times less prototypical. This fuzziness requires a different method of representation—some mechanism for indicating relative weight and presence of attributes and relative closeness or distance from some best-example prototype. With permeable membranes and dynamic membership in classes, it is difficult to maintain the principles of transitivity and inheritance.

In summary, hierarchies are excellent representations for knowledge in mature domains in which the nature of the entities, and the nature of meaningful relationships, is known. Hierarchies are useful for entities that are well defined and have clear class boundaries. In general, some theory or model is necessary to guide the identification of entities, the rules of association and distinction, and the order in which these rules are invoked:

Trees. Another type of classificatory structure used to represent entities and their relationships is a tree. A tree divides and subdivides its classes based on specific rules for distinction just as in a hierarchy but does not assume the rules of inheritance. Thus, in a tree, the entities have systematic relationships but not the generic (is-a) relationship. There are many types of relationships that can be represented by a tree (see, for example, Figure 2).

GENERALS
COLONELS
CAPTAINS
LIEUTENANTS
SERGEANTS
PRIVATES

Figure 2. Tree: Chain of Command in the Army.

In this tree, the entities are the names of Army ranks. The relationships among the ranks can be described as "chain of command" or "who reports to whom." That is, a GENERAL *commands* COLONELS, COLONELS *report to* GENERALS, COLONELS *command* CAPTAINS, and so on. GENERALS *command* PRIVATES as well, although not directly, but a PRIVATE is not a kind of SERGEANT, and a SERGEANT is not a kind of LIEUTENANT, so the principle of division by *species/differentia* does not apply. Conversely, SERGEANTS do not inherit the attributes of LIEUTENANTS. In terms of knowledge representation, a tree works well to display a particular relationship and the distribution of the entities vis-à-vis that relationship. This tree shows who is on top and who is on the bottom of the chain of command. Some inferences can be made about prerogatives and responsibility, but only weakly since these inferences are based on pragmatic knowledge and not on knowledge that is stored in the structure of the classification itself. By knowing something about the domain, it is also possible to guess that GENERALS once were PRIVATES and thus bring "up the ladder" all of the experiences of going through the ranks, but this is not a formal requirement of the representation either, and may, in fact, be wrong.

Furthermore, a tree is "flatter" in its representation than is a hierarchy; there is less richness in the representation at each level because there is no inheritance or sharing of attributes. For example, there is no indication of the nature of LIEUTENANTS—their essence as it were—from their position in this classification. In a hierarchy, if we know a *dog* is a *mammal*, we know something about the mammalian attributes of the entity *dog*. What are the attributes of a LIEUTENANT that we can learn from the classification? Does a LIEUTENANT share attributes with a GENERAL but has less of them or different kinds? This type of information is not included explicitly. All we can know from this tree is that one rank commands the one lower in the pecking order.

Another kind of tree is one in which the entities are related by the partitive relationship. This means that each class is divided into its components, these components into subcomponents, and so on (see, for example, Figure 3).

In this example, SYRACUSE is *part of* ONONDAGA COUNTY, which in turn is *part of* NEW YORK, and so on. The partitive relationship (also known as *part/whole*) is a richer representation than the one shown in Figure 2. This is because the principle of inclusion allows more information to be shared. For instance, SYRACUSE is not *a kind of* ONONDAGA COUNTY, so what is true of the county *as a county* is not true for the city in it, but SYRACUSE is *part of* ONONDAGA COUNTY and therefore inherits those attributes of the county that pertain to all units within it (e.g., location in New York State, climate, and so on). This relationship is so rich in representational power, in fact, that in many classification schemes there is no distinction made between the purely hierarchical and partitive

NORTH AMERICA
- CANADA
  - ALBERTA
  - +
  - +
- MEXICO
  - +
  - CHIAPAS
  - +
- UNITED STATES
  - ALABAMA
  - +
  - +
  - NEW YORK
    - ONONDAGA COUNTY
      - SYRACUSE
  - +
  - +
  - WYOMING

Figure 3. Part/Whole Relationship.

relationships, and many people refer to both as "hierarchies." There is some psychological support for this, since both pure hierarchies and part/whole classifications convey the notion of going from the more general and inclusive to the more specific or elemental.

Care must be taken, however, in making use of tree representations to ensure that the correct attributes are drawn upon in making inferences. This problem becomes clearer in another part/whole example (see Figure 4).

AUTOMOBILE
- BODY
- ENGINE BLOCK
  - PISTONS
  - VALVES
- INTERIOR
  - +
  - +

Figure 4. Part/Whole Relationship.

VALVES are *part of* the ENGINE BLOCK, but the nature of VALVES is distinct from the nature of PISTONS, and it would be incorrect to assume (despite their sibling position in the classification) that they share many attributes in common the way wolves and dogs do. In fact, VALVES and PISTONS are not similar entities at all. They share the attribute of being part of the ENGINE BLOCK, but that is only a partial explanation of what they are—

their essence. It would not be sufficient knowledge for most practical purposes the way knowing that dogs and wolves are closely related might prove useful. So, trees have the following formal requirements:

- ***Complete and comprehensive information.*** Just like in a hierarchy, the entities that will be included in a tree must be decided in advance. First, it must be decided what will constitute an entity. Knowledge about the entities must be relatively complete in order to decide on the scope of the classification and the important criteria of distinction.
- ***Systematic and predictable rules for distinction.*** The general structure of a tree is determined by the relationships among the entities. Part/whole relationships might be appropriate for some knowledge, while other relationships (such as cause/effect; starting point/outcome; process/product; and so on) might be appropriate for other types of knowledge representation. These relationships should be ones that best reveal the knowledge of the domain—that is, the way in which all the entities interact with each other.
- ***Citation order.*** In both hierarchies and trees it is important to decide the order in which rules of distinction will be invoked. The most important of these decisions is the "first cut" because this determines the shape and eventually the representational eloquence of the classification. If the first cut is a trivial one, the rest of the tree becomes awkward and does not reflect knowledge very well. For example, in the biological classification of animals (a hierarchy), the first cut is: *has a backbone/does not have a backbone (vertebrate/invertebrate).* While this cut produces a very skewed distribution in terms of numbers of species (there are many times more invertebrate species than vertebrates), the resulting classification proceeds smoothly down the subdivisions and is able to cluster many attributes that "make sense" with respect to what we know about fundamental qualities of animals. In trees, the determination of an appropriate citation order is all the more important because trees are essentially descriptive, and the picture they present will depend on the first branching. For instance, in the AUTOMOBILE example presented above, it would be possible to make the first division BACK OF CAR/FRONT OF CAR/MIDDLE OF CAR, and proceed to decompose those sections into their component parts. But would this make sense? Would it present a reasonable division of an automobile's components? Would it help us with knowledge about cars? Perhaps for someone in some context. There is no easy answer to what constitutes a meaningful division, and the decision often rests on consensual models or tradition.

Trees are useful knowledge representations for the following reasons:

- ***Highlight/Display relationship of interest.*** This is the primary strength of a tree. It lays out the entities comprising a domain in a pattern of classes that highlights or makes evident the important or defining relationships among them.
- ***Distance.*** A tree reveals the distances between entities (either physical distances or metaphorical ones). Thus one can determine that a COLONEL is "closer" to a GENERAL than is a PRIVATE, at least along the dimension of chain of command. If entities are components of the same super component, this means they are "closer" in space or in function.
- ***Relative frequency of entities.*** This feature of trees is also shared by hierarchies. When entities cluster in large numbers under one classification label, this is frequently an opportunity for the creation or discovery of new rules for distinguishing among them. When a cluster is small and has only a few entities, these entities tend to be treated as if they were all the same. It may be neither feasible nor reasonable to make distinctions among them, and taking account of any differences may not support the enterprise. Once the cluster grows, however, and the number of entities reaches a critical mass, it might be useful to further differentiate them. In such a case it is necessary to discover new knowledge that will suggest the best way of making these finer distinctions. Conversely, when a category consistently has a member of one or just a few, it might signal the need for merging categories and rethinking the logic behind the division in the first place. In this case also, it is necessary to generate new knowledge in order to guide the merging or shifting of the orphan categories.

The use of trees as knowledge representations shares some of the same problems as does the use of hierarchies:

- ***Rigidity.*** Because a tree is characterized by the relationships among entities and the citation order, the general shape of the tree—its expressiveness as a knowledge representation—is determined a priori. This means that new entities can be added, if they fit into a place in the structure but, if the new entity or new knowledge does not fit well, the entire structure must be rethought and sometimes rebuilt.
- ***One-way flow of information.*** In a hierarchy, information flows in two directions: vertically, between classes, superclasses, and subclasses, and also laterally, between sibling classes (classes sharing the same superclass). In a tree, even if it is a part/whole representation, the information flows in a vertical direction up and down. Siblings in a class may in fact be entirely different types of objects. So there are rules for *species* but not for *differentia.* Many people assume that, since Syracuse is in New York State and New York City is also in New York State, that they are similar when in fact they only share the attribute of being in

the same state and little else. Syracuse may be more like some other city in another state than it is like New York City. At any rate, the tree classification is not particularly good at representing multidirectional complex relationships.

- ***Selective perspective.*** As with hierarchies, by emphasizing a certain relationship, a tree can mask, or fail to reveal, other equally interesting relationships. For instance, in the Army ranks example, the only relationship available to us is the "who commands whom" relationship. It does not touch upon the relationship of ranks when in combat, for instance, as opposed to the relationships among ranks of troops stationed at home. It does not show the distribution of men to women in the various ranks, or the distribution of ethnic or racial groups, and so on. It is completely silent on the classification of functional jobs in the Army (such as nurses, quartermasters, and so on). In other words, there are many other perspectives or lenses through which one could "know" the Army. The typology of ranks based on who commands whom is but one of them.

In summary, trees are useful for displaying information about entities and their relationships along one dimension of interest. They require fairly complete knowledge about a domain or at least about one aspect of a domain. A tree representation is good for displaying the relative placement of entities with respect to each other and their frequency at any node. On the other hand, trees are limited in how much they can represent, especially in terms of knowledge about entities within the same class. Furthermore, trees allow only partial inference.

Paradigms. A third classificatory structure is one in which entities are described by the intersection of two attributes at a time. The resulting matrix (or paradigm) reveals the presence or absence and the nature of the entity at the intersection (see Figure 5).

In this representation, we see two axes. The vertical has headings designating gender; the horizontal, types of kinship relationships. The cells represent the labels or names for this intersection of gender and kinship relationship. In this example, we have combined two such paradigms: one in English and one in Polish. You could imagine each one standing on its own but, for purposes of comparison, we have superimposed one on the other. Paradigms have the following formal requirements:

- ***Two-way hierarchical relationship.*** Each cell entity (in this case the label signifying a kinship relationship) is related to both the vertical and the horizontal axis by a generic relationship. For instance, a *Mother is a Female,* and a *Mother is a Parent.* Across the row, all the entities are related to each other in being a subclass of the row header. So, *fathers,*

Kinship Relationship								
	Parent		*Sibling*		*Parent's Sibling*		*Parent's Sibling's Child*	
	Eng.	Pol.	Eng.	Pol.	Eng.	Pol.	Eng.	Pol.
Male	Father	Ojciec	Brother	Brat	Uncle	Stryj (father's side) Wujek (mother's side)	Cousin	Brat Stryjeczny (father's side) Brat Cioteczny (mother's side)
Female	Mother	Matka	Sister	Siostra	Aunt	Stryjenka (father's side) Ciocia (mother's side)	Cousin	Siostra Stryjeczna (father's side) Siostra Cioteczna (mother's side)

Figure 5. A Paradigm Displaying a Selection of KinshipTerms in English and Polish.

brothers, uncles, cousins, and *fathers-in-law* are all *males.* Entities in columns are related to each other by being in the same subclass. So, *uncles* and *aunts* are both *siblings of parents.* There is a shallow hierarchy running in both directions. However, entities are not related to each other in a generic relationship. Thus a *mother* is not a kind of *father* nor do they inherit properties from each other.

- ***Axes represent two attributes of interest.*** Each axis represents one attribute that might serve to describe the entities in a meaningful way. In the example, the two axes represent "*the sex of the person*" and "*the way the person is related.*" The interesting feature of a paradigm is that it affords us a view of the entities classified along two dimensions at once.
- ***Cells may be empty or may have more than one entity.*** Paradigms not only show us the intersection of two attributes, but also show us the presence, absence, and frequency of entities at these intersections.

So, how do we use this classificatory structure to represent and create knowledge?

- ***Naming.*** Paradigms are frequently used in the study of terminology. As mentioned in a previous section, hierarchies enable the creation of strong definitions, but paradigms allow the study of patterns of naming. When people name things, they are creating an abstraction by

incorporating a complex set of attributes under one label. Objects that are quite different in many ways but share defining attributes may still be given the same name. For instance, animals with a wide range of physical attributes are labeled *dog* if they share the defining attributes. Or, when we call something a *hamburger,* we may include under this rubric many slightly different kinds of sandwiches. They may have lettuce, or a slice of onion, or ketchup, or not; they may be small or large, but if they have a beef patty and a bun they are named *hamburger.* Now, if you add a slice of cheese, the name changes. Two hamburgers that are quite similar with respect to lettuce, tomato, onion, ketchup, and even sesame seeds, will still be named differently if one has a slice of American cheese. Naming will vary according to context, region, profession, and so on. So terminology indicates classificatory decisions, and paradigms serve as descriptive displays of terms as well as tools for analysis.

- ***Distinction and lack of distinction.*** Paradigms can show the extent to which the intersecting criteria have distinct terms. In our example, we see that English has a single label for all the relationships displayed, while Polish has two terms each for uncles and aunts, and four terms for cousins, depending on whether they are related through the mother or father. So, in English, there is no distinction at all between cousins, and only a gender distinction between labels for your parents' siblings. Furthermore, besides being distinguished by side of the family, *cousins* in Polish are not completely distinguished from siblings and are given names that have the same root term as do *brothers* and *sisters.*
- ***Patterns of similarity and difference.*** In terms of knowledge creation, paradigms often provide a heuristic tool for the discovery of regularity in the patterns of distinction. When distinctions are made in naming (that is, when people create different labels for concepts), we assume that the criterion for having made that distinction is important in some way. In this case, the distinction of relationship through either the mother's or father's side is important in Polish. Conversely, while English has distinct terms for cousins and siblings, Polish uses similar terms for both, distinguishing only by gender. The knowledge conveyed is that, even though each culture has a great deal of overlap in equivalent terminology, there are some subtle differences that may have historical or other explanations. In fact, it is interesting to note that the Polish distinction between aunts and uncles from different sides is fading. Did English once have such distinctions? Does this indicate the cultures are merging?
- ***Empty cells.*** Empty cells in a paradigm provide an opportunity to investigate the reasons for the lack of a term. Does the absence of a term indicate the absence of a concept or does it indicate that the criteria chosen for the axes are not meaningful ones? Does every language,

for instance, have the notion of a "kissing cousin?" Why is there no female equivalent of *misogynist?*

The limitations of paradigms as knowledge representation and discovery tools are as follows:

- ***Requires knowledge of domain.*** The expressiveness of a paradigm relies on the felicitous choice of the attributes represented on the two axes. The fidelity of the picture that a paradigm reveals can be compromised if the dimensions are trivial and do not reflect fundamental concepts. In our example, the axes are chosen from concepts well established in the field of cultural anthropology: kinship expressed through blood and marriage, as well as distinctions made by gender. Paradigms that use dimensions guided by theory or a model usually do a better job of reflecting knowledge in the domain because they rest on a consensual framework of description. In other words, they are using a common vocabulary for communication. In fields where the fundamental relationships or concepts are not well understood, it is difficult to build a paradigm that reveals essential knowledge.
- ***Limited perspective.*** While a well-chosen set of dimensions may produce a valid description, it also produces a filter that limits the scope of what might be seen. So, in the example, kinship terms are expressed using blood/marriage relationships and gender as dimensions of distinction. To us these distinctions seem self-explanatory and almost universal, but in fact they are artifacts of our own cultural assumptions that we then impose on our observations of the world. Consider, for instance, that the dimensions do not address other family bonds, such as those that are based on strong affinity, legal adoption, and other socially invented forms of kinship. Nor do they allow for other cultural definitions of the entities themselves, such as alternative views of what constitutes a *parent.* Thus, the paradigm presented in the example is a view through a particular lens. Another set of dimensions would present a different view and would most likely produce different analytical outcomes.
- ***Limited explanatory power.*** Because paradigms invoke dimensions only in pairs, they (like most classificatory structures) rarely produce a complete picture of a phenomenon. While paradigms do use the potentially rich representation of a hierarchical relationship vertically and horizontally, in a paradigm this relationship is shallow—only one deep—and therefore not very complex. For this reason, paradigms are essentially descriptive. They help clarify; they may *suggest* patterns and anomalies, but these patterns are not inherent to the structure and must be interpreted by the person using the paradigm.

In summary, paradigms are good tools for discovery. They reveal the presence or absence of names for entities defined by pairs of attributes. They

can be used for comparison and for the display of patterns and anomalies with respect to the variety and distribution of terms. Paradigms are heuristic in that they present a clear view that can then be analyzed and interpreted. Like most classificatory structures, paradigms require knowledge of the domain or some guiding principles in order to make a good choice of dimensions and, like most classificatory structures, paradigms are usually partial and biased representations.

Faceted Analysis. Faceted classifications are not really a different representational structure but rather a different approach to the classification process. The notion of facets rests on the belief that there is more than one way to view the world, and that even those classifications that are viewed as stable are in fact provisional and dynamic. The challenge is to build classifications that are flexible and can accommodate new phenomena.

Faceted classification has its roots in the works of S.R. Ranganathan, an Indian scholar, who posited that any complex entity could be viewed from a number of perspectives or *facets.* He suggests that these fundamental categories are Personality, Matter, Energy, Space, and Time (Ranganathan, 1967). Over the years, Ranganathan's facets have been reinterpreted in many contexts, but it is surprising how well they have weathered the test of time. They have been used to classify objects as disparate as computer software (for reuse), patents, books, and art objects (Kwasnik, 1992a).

Not all faceted classifications use Ranganathan's prescribed fundamental categories, but what they do have in common is the process of analysis.

Period/Style	*Place*	*Process*	*Material*	*Object*
19th Century	Japanese	raku	ceramic	vase
Arts & Crafts	American		oak	desk

Figure 6. A Faceted Analysis of Artifacts.

Figure 6 shows a possible solution to the classification of material culture which, in its diversity, defies easy description and categorization. For purposes of demonstration, this is a simplified version of the one used by the *Art and Architecture Thesaurus.* For any given artifact, there are many possible ways of representing it, let alone the "knowledge" that enabled its production or its value. The faceted approach follows these steps:

- ***Choose facets.*** Decide, in advance, on the important criteria for description. These form the facets or fundamental categories. In this case we have Period, Place, Process, Material, and Object, following closely on what Ranganathan suggested.
- ***Develop facets.*** Each facet can be developed/expanded using its own

logic and warrant and its own classificatory structure. For example, the Period facet can be developed as a timeline; the Materials facet can be a hierarchy; the Place facet a part/whole tree, and so on.

- ***Analyze entities using the facets.*** In analyzing an entity, choose descriptors from the appropriate facets to form a string, as shown above. Thus, the classification string for object 1 is "19th Century Japanese raku ceramic vase." The string for object 2 is "Arts & Crafts American oak desk." It is important to note that the process is not one of *division* (as in a hierarchy) where the entities are subdivided into ever more specifically differentiated categories. It is not a process of *decomposition* either (as in a part/whole tree), in which the entities are broken down into component parts, each part different from the whole. Instead, the process of *analysis* is to view the object from all its angles—same object but seen from different perspectives. So, in the example, the vase can be seen from the point of view of its period, the place in which it was made, the material and processes, and so on.
- ***Develop citation order.*** In organizing the classified objects, choose a primary facet that will determine the main attribute and a citation order for the other facets. This step is not required and applies only in those situations where a physical (rather than a purely intellectual) organization is desired.

The development of a faceted approach has been a great boon to classification because it meshes well with our modern sensibilities about how the world is organized. Specifically, it is a useful tool because it:

- ***Does not require complete knowledge.*** In building a faceted scheme, it is not necessary to know either the full extent of the entities to be accommodated by the scheme nor the full extent of the relationships among the facets. It is thus particularly useful in new and emerging fields or in fields that are changing.
- ***Hospitable.*** When a classification is hospitable it means it can accommodate new entities smoothly. In a faceted scheme, if the fundamental categories are sound, new entities can be described and added. This is particularly important in the classification of objects such as cultural artifacts, where we have no way of predicting the things that will be produced by the human imagination. If an artifact produced 100 years from now could be described by the fundamental categories of period, place, material, process, and object, then the classification scheme will still be robust.
- ***Flexibility.*** Since a faceted scheme describes each object by a number of *independent* attributes, these attributes can be invoked in an endlessly flexible way in a sort of Lego approach. "Let me see all the iron objects made in 17th-century Scotland." "OK, now all the copper objects." "OK, now iron objects in Italy...". This flexibility can be used to

discover new and interesting associations. The approach is called *post-coordination* and means that attributes can be mixed and matched at the time of retrieval. It is in contrast to the *pre-coordinated* categories that are a requirement of most hierarchies in which the rules for class inclusion are invoked at the time the entity is classified and stay fixed from there on.

- ***Expressiveness.*** A faceted approach can be more expressive because each facet is free to incorporate the vocabulary and structure that best suits the knowledge represented by that facet.
- ***Does not require a strong theory.*** Since a faceted classification does not have an overall structure, it does not have to have a "theoretical glue" to hold it together and to guide the rules for association and distinction. It can be constructed ad hoc so long as the fundamental categories function well.
- ***Can accommodate a variety of theoretical structures and models.*** A faceted approach makes it possible to represent a variety of perspectives as well. For instance, in facet analyzing a piece of literature, one facet may reflect a particular model of genres, another a model of languages, and so on. In a traditional hierarchy, it may be extremely difficult or impossible to blend the two, while a faceted scheme allows them to co-exist.
- ***Multiple perspectives.*** One of the most useful features of a faceted approach is that it allows entities to be viewed from a variety of perspectives—a feature that is lacking in hierarchies and trees. In a faceted analysis, it is possible to describe a *dog* as an *animal*, as a *pet*, as *food*, as a *commodity*, and ad infinitum, so long as the fundamental categories have been established with which to do this.

While the flexibility and pragmatic appeal of faceted classifications have made this a popular approach, there are some limitations in terms of knowledge representation and creation:

- ***Difficulty of establishing appropriate facets.*** The strength of a faceted classification lies in the fundamental categories, which should express the important attributes of the entities being classified. Without knowledge of the domain and of the potential users, this is often difficult to do. While it is possible to flexibly add entities, it is not a simple matter to add fundamental facets once the general classification is established.
- ***Lack of relationships among facets.*** Most faceted classifications do not do a good job of connecting the various facets in any meaningful way. Each facet functions as a separate kingdom, as it were, without much guidance as to how to put the parts together. For example, to facet analyze motion pictures by *genre, country, director, film process*, and so on, we would still have no insight as to the meaningful relationships of, say, a particular country and the popular film genre there or of a

particular film process and the genres it supports. In terms of theorizing and model building, the faceted classification serves as a useful and multidimensional description but does not explicitly connect this description in an explanatory framework.

- ***Difficulty of visualization.*** A hierarchy or a tree, and especially a paradigm, can be visually displayed in such a way that the entities and their relationships are made evident. This is difficult to do for a faceted classification, especially if each facet is structured using a different internal logic. As a result, faceted schemes can only be viewed along one or two dimensions at a time, even though a more complex representation is actually incorporated into the descriptive strings. Thus it is difficult to see a vase in the context of other vases, of other Japanese artifacts, of other clay objects, of other raku objects, and so on, *all at the same time.*

Nevertheless, faceted schemes continue to flourish because we recognize that they allow at least some systematic way of viewing the world without the necessity for a mature and stable internal framework in which to view it. Information technology has promise for new ways of enabling multidimensional visualization and for developing computer-assisted ways of discovering patterns and anomalies that can possibly lead to new knowledge.

Classification and Knowledge

There are many ways in which classification schemes and knowledge interact. Sometimes the interaction is so harmonious that the two remain linked for a long time. Sometimes knowledge changes and the classification must also change or knowledge changes and the classification is no longer adequate to the task. Sometimes the classification itself generates new knowledge. The following discussion is representative of ways in which knowledge and classifications mutually interact.

Changing Explanatory Frameworks

The Periodic Table of Elements, attributed to Mendeleyev, is an example of a classification scheme that has endured through several explanatory frameworks. When the Periodic Table was first proposed, there was already a body of knowledge about individual elements—i.e., facts and observations, including the knowledge of atomic weight. It was observed that elements could be arranged in a systematic order according to atomic weight, and this would show a periodic change of properties. This early Periodic Table proved to be a very useful tool, leading to the discovery of new elements and a new understanding of already known elements. In terms of theory, the Table "so divide[d] its subject matter that it [could] enter into many and important true propositions about the subject matter..." (Kaplan, 1963, p. 50). The rule for determining one element from another was atomic weight, a basically descriptive criterion, but it was not

until the discovery of atomic theory that the periodicity of the table was fully understood. This theory *explained* (rather than merely described) the underlying principles behind the regularity and pattern of the classified entities. With this new explanation, many new properties could be inferred. The table became a predictive tool for as yet undiscovered elements as well as an explanatory tool and a very fruitful descriptive tool. It reflected well what was already known about elements and pointed to new knowledge (such as the common characteristics of inert gases). What is interesting is that the original Periodic Table did not have to undergo fundamental changes in structure even though a new explanatory framework was discovered.

Changes in Perspective

Technological advances in measuring and viewing instruments have had a profound influence on classifications. This is because new instruments reveal new knowledge that does not always fit neatly into existing knowledge representation structures. Such instruments have included carbon dating, the electron microscope, DNA testing, remote sensing, and so on. For instance, clouds were traditionally very simply classified by shape and by height from the horizon. This classification was developed when we could only see clouds from our perspective standing on the earth. Now we can measure the moisture, temperature, particulate matter, and charge of a cloud. Moreover, clouds can be observed from a satellite thereby observing global patterns. We know about the typical life cycle of a cloud—how clouds change shape and identity. This new understanding of clouds has a profound impact on weather forecasting, navigation, and other fields of knowledge, yet the traditional classification is robust as a simple and clear form of communication. It remains very popular and coexists with new classifications.

Sometimes the new way of observation yields a new classification. For instance, gems used to be classified on a scale of hardness (with diamonds at one end and chalk at the other) and also by color. These attributes were visible to the naked eye. Once it was possible to view gems through a microscope, it was then necessary, or at least more useful, to alter this classification to include knowledge about crystalline structure.

Changing Entities

Sometimes, though, changes in the way we can observe have led to fundamental changes not only in the classificatory structures but also in the nature of the entities themselves. For instance, a complex problem exists in trying to coordinate traditional ways of describing natural habitats with new ways of observation and measurement. Formerly, scientists spoke in terms of rainfall, temperature, growth forms, dominant life forms, and so on. These attributes were described and classified according to whatever model predominated or was accepted and resulted in such

constructs as *deserts, tropical rainforests,* and so on. But today the mapping of natural habitats is done by remote sensing data that measure a different kind of unit such as reflectance, texture, density, spatial patterns, slope, Leaf Area Index, and so on (Muchoney, 1994)—that is, each of these parameters does not necessarily correspond to something (an entity) that we can unambiguously call "a tree" or a "a camel." Remote sensing data do not require a semantically coherent entity to be the cause of measurement. All that matters is that measurements can be taken and this measurement can then be combined, clustered, and analyzed with other measurements to yield "types." Put another way, the resulting entities are not semantically meaningful in the same way as traditional names. The problem arises in communication about the habitats and also in theorizing about them—i.e., in making sense of the phenomena.

Another example of changing entities within a domain has occurred in the classification of musical instruments. Instruments are classified basically by material and the method of producing sound (striking, blowing, bowing, strumming, and so on). This classification did quite well with a few slight adjustments here and there for hybrid instruments, but it hit a real snag with the introduction of electronic instruments such as synthesizers. The problem lies not so much in squeezing the newcomer categories into the old scheme but rather in how well the old criteria fit the new entities as meaningful guides for association and discrimination. This situation is a classic example of shifts that occur only occasionally at first but might eventually lead to a complete overhaul of the classification or perhaps to the creation of two parallel classifications.

Classifications Built When There is No Consensus

Many classifications must be built when there is no generally accepted theory or model on which to construct or to define the entities. For example, the classification of mental disorders is mandated economically by insurance companies and legal requirements. Social institutions require that we be able to determine who is legally sane, who must be confined to care, and who will be reimbursed for services. This classification is therefore built on factors that are not based on any particular theory of mental illness or mental processes but rather on readily observable symptoms and behaviors. It is therefore only a moderately good descriptive tool and an almost useless tool for understanding fundamental processes from any theoretically coherent perspective. Thus, the classification fails to act as a heuristic device by generating provocative questions or providing interesting insights.

Classifications Where There are No Uniform or Stable Entities

Scientists have long been struggling with a classification of smells. While we are able to construct useful classifications of colors based on what we know about the physics of light, the psychology and aesthetics of

color and the human ability to perceive and use colors, we have not had the same success with smells. We are forced to refer to smells using analogies: *fruity, citrus, green, floral, putrid,* and so on. One of the problems is that there is no "unit" of smell, no building block that could then be classed and differentiated in some systematic way. This lack of a classification has led to reliance on the essentially subjective artistry of individuals with respect to the identification, blending, and general understanding of smell. This does not mean that we do not know how smell works in terms of perception. Nor does it mean that we do not understand the powerful nature of smell in human life, but the fact remains that we have no good way of talking about smells as smells.

A similar problem arises in the classification of viruses. This is not because viruses do not exist as entities but, rather, because they change and are never in a form unambiguous enough to be pinned down by a clear classification.

The Intersection of Theory and Economic Interests

Classifications are never created in a political or social vacuum. Everyone is familiar with the old Department of Agriculture classification of food groups: meats, dairy and eggs, grains, fruits and vegetables, and fats. From a classification point of view, the dairy and eggs category always seemed a rather odd one in that it is not clear on what basis dairy products and eggs go together, nor along which dimension they are distinguished from the other categories. It is not by source (animals) nor by nutritional component (protein). Furthermore, the classification does not indicate the relative importance of each group. "At least one from each" was the slogan. As it turns out, the classification was the result of a strong lobby by the dairy industry. It is amazing how well established this classification became and how long it persisted. The new classification is really quite elegant. It builds a pyramid with grains on the bottom and fats on the top. Dairy and eggs now share a level with meats. This new classification reflects modern nutritional science much more coherently. It not only classes the foods according to some understandable criteria, but also indicates, by the narrowing pyramidal shape, the relative amount of foods from each of the groups that should be consumed. This classification may, as it turns out some day, reflect faulty scientific knowledge, but at least it reflects it with fidelity and clarity.

Keys and Other "Thin" Classifications

A key is a classification that is built using an easily identifiable, but not necessarily theoretical, set of criteria. One example is a field guide to flowers. In such a guide, flowers are classed first by their petal color. Petal color is a characteristic that is easy to identify but is trivial when compared to more fundamentally meaningful attributes of a flower such as plant structure or reproductive mechanisms. The petal color yields relatively

little fundamental knowledge of the flower, but petals are an easy way to narrow the field of possibilities, especially for the novice. A key, therefore, allows easy entry into a deeper classification. However, consider the classification of baseball in the Dewey Decimal Classification in Figure 7.

796	Athletic and outdoor sports and games
796.3	Ball games
796.35	Ball driven by club, mallet, bat
796.357	***Baseball***

Figure 7. Dewey Decimal Classification of Baseball.

This scheme positions baseball, along with other ball games in which a ball is hit with a mallet, club, or bat, right next to field hockey and croquet as well as polo. This is not an inaccurate classification. There are no factual errors, as it were, but by using the ball and the bat as the defining criteria, the classification is reduced to a very thin representation of what baseball is. It does not address any of the team aspects; the cultural aspects; or the aesthetic, athletic, economic, strategic, or spiritual aspects of baseball as a sport. Why? Because this is difficult to do, and there is no consensus really of what these attributes might be or how they might be expressed. There is no generally accepted theory of games or sports, but we can all readily agree that yes, indeed, in baseball you hit a ball with a bat. In this case, the classification is structured like a key but does not lead to a deeper theoretical representation of baseball in all its complexity.

Conclusion

Classification is a way of seeing. Phenomena of interest are represented in a context of relationships that, at their best, function as theories by providing description, explanation, prediction, heuristics, and the generation of new questions. Classifications can be complex or simple, loaded with information or rather stingy in what they can reveal. They can reflect knowledge elegantly and parsimoniously, or they can obfuscate and hinder understanding. Some classifications enable flexible manipulation of knowledge for the purposes of discovery; some are rigid and brittle, barely able to stand up under the weight of new knowledge. It is useful to understand the properties of various classification structures so we can exploit their strengths and work around the weaknesses. In the future, classification will be enhanced by new methods of revealing patterns, associations, and structures of knowledge, and by new ways of visualizing them.

NOTES

[1] Some of the ideas in this article were first described in a paper presented at the Third ASIS SIG/CR Workshop on Classification Research, Pittsburgh, PA, 1992 (Kwasnik, 1992b).

[2] Adapted from an example in Aitchison, Jean and Gilchrist, Alan (eds.). (1987). *Thesaurus construction*, 2d ed. London, England: Aslib, p. 80.

REFERENCES

Ackrill, J. L. (Trans.). (1963). *Aristotle's "Categories" and "De Interpretatione": Translated with notes.* Oxford, England: Oxford University Press.

Bronowski, J. (1978). *The origins of knowledge and imagination.* New Haven, CT: Yale University Press.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation, 45*(4), 273-301.

Kwasnik, B. H. (1992a). The legacy of facet analysis. In R. N. Sharma (Ed.), *S.R. Ranganathan and the West* (pp. 98-111). New Delhi, India: Sterling.

Kwasnik, B. H. (1992b). The role of classification structures in reflecting and building theory. In R. Fidel, B. H. Kwasnik, & P. J. Smith (Eds.), *Advances in classification research, vol. 3* (Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop) (pp. 63-81). Medford, NJ: Learned Information, for the American Society for Information Science.

Muchoney, D. M. (1996). Relationships and divergence of vegetation and mapping classifications. In R. Fidel, C. Beghtol, B. H. Kwasnik, & P. J. Smith (Eds.), *Advances in classification research, vol. 5* (Proceedings of the 5th ASIS SIG/CR Classification Research Workshop). Medford, NJ: Learned Information, for the American Society for Information Science.

Olson, H. A. (1998). Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library Trends, 47*(2), 233-254.

Ranganathan, S. R. (1967). *Prolegomena to library classification,* 3d ed. Bombay: Asia Publishing House.

Implicit Text Linkages between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery

DON R. SWANSON AND NEIL R. SMALHEISER

ABSTRACT

THE PROBLEM OF HOW TO FIND INTERESTING but previously unknown implicit information within the scientific literature is addressed. Useful information can go unnoticed by anyone, even its creators, if it can be inferred only by considering together two (or more) separate articles neither of which cites the other and which have no authors in common. The two articles (or two sets of articles) are in that case said to be complementary and noninteractive. During the past twelve years, this project has uncovered and reported numerous complementary relationships in the biomedical literature that have led to new information of scientific interest. Several of these literature-based discoveries subsequently have been corroborated through clinical or laboratory investigations. We describe how to use software that can create suggestive juxtapositions of Medline records, the purpose being to help biomedical researchers detect new and useful relationships. This software, called Arrowsmith, has also proved valuable as a tool for investigating patterns of complementary relationships in natural language text (Arrowsmith can be used free of charge at http://kiwi.uchicago.edu).

INTRODUCTION

The juxtaposition of certain natural language text passages from different biomedical journal articles can reveal or suggest new information not contained in the original passages considered separately. For example,

Don R. Swanson, Division of the Humanities, The University of Chicago, 1010 E. 59th St., Chicago, IL 60637
Neil R. Smalheiser, Department of Psychiatry, University of Illinois, 1601 W. Taylor St., Chicago, IL 60612
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 48-59

one article might report an association or link between substance A and some physiological parameter or property B while another reports a relationship between B and disease C. If nothing has been published concerning a link between A and C via B, then to bring together the separate articles on A-B and B-C may suggest a novel A-C relationship of scientific interest. There are now about 9 million records in the Medline database, and hence about 40 trillion (40,000,000,000,000) possible pairings of records. Clearly the vast majority of record pairs and article pairs have never been considered together. It is plausible to think that there are many undiscovered implicit relationships within the biomedical literature, at least some of which might be important (Swanson, 1993, pp. 611-19). It is important, therefore, to develop systematic methods for finding them.

The possibility of literature-based discovery implied by the above model underscores two important properties of sets of scientific articles—complementarity and noninteractivity. Two sets of articles are defined here as complementary if together they can reveal useful information not apparent in the two sets considered separately; two sets are defined as noninteractive if they are disjoint and if no article in either set cites, or is co-cited with, any member of the other set (Swanson, 1987, 1990a, 1991).

The first three examples of "undiscovered public knowledge" (Swanson, 1986a, 1986b, 1988, 1990c) demonstrated that complementary noninteractive structures actually do exist within the biomedical literature and can lead to the discovery of apparently new and interesting implicit relationships. In at least two of these cases (Swanson, 1986a, 1988) the hypothesis was subsequently corroborated experimentally by medical researchers. We have cited and discussed these corroborations elsewhere (Swanson, 1993; Smalheiser & Swanson, 1994). The hypothesis advanced in Swanson (1990c)—that the anabolic effects of arginine are brought about by systemic or local release of somatomedin C—has also received direct supporting evidence in three recent studies (see Kirk, 1993; Hurson, 1995; Chevalley, 1998); a fourth study by Corpas (1993) reported negative results. Gordon and Lindsay (1996) re-examined, replicated, and extended Swanson's work (1986a).

The above structures were found through innovative, partially systematic, database search strategies (Swanson, 1989a, 1989b). Computer-assisted processing of the downloaded output enhanced the user's ability to discover novel implicit relationships (Swanson, 1991). This software evolved into a system called Arrowsmith that processes article records downloaded from large bibliographic databases such as Medline. Text passages within database records provide the raw material that suggests or points to underlying linkages (such as A-B and B-C above) between separately published scientific findings or arguments. Our goal has been to create a research tool for studying complementary noninteractive structures in the scientific literature and at the same time to create a working system useful

to biomedical scientists (Swanson, 1991; Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1998b).

With the help of Arrowsmith, we have developed five additional examples of complementary noninteractive literature structures (Swanson & Smalheiser, 1997; Smalheiser & Swanson, 1994, 1996a, 1996b, 1998a), each of which led to a novel, plausible, and testable medical hypothesis. One of these studies (Smalheiser & Swanson, 1998a) elicited publication of a concurring letter from an author whose work was the basis for a new hypothesis that we proposed (Ross, 1998).

The Process of Inferring Text Linkages

Given two Medline titles that appear to be linked, the process of inferring a biologically meaningful linkage may be more subtle than it seems at first sight. We consider here examples taken from Swanson (1988):

1. "The Relation of Migraine and Epilepsy" (p. 551)
2. "Preliminary Report: The Magnesium-Deficient Rat as a Model of Epilepsy" (p. 556).

The two titles taken together appear to provide a link, via epilepsy, between migraine and magnesium deficiency (epilepsy being just one of the eleven links reported). The role of Arrowsmith in this example is only to bring the two titles together in order to create a suggestive juxtaposition. Whether the relationship thus revealed might merit further investigation then depends on human judgment. Such judgment in general would be difficult to replace by a computer procedure, for it almost inevitably entails certain background knowledge, context, and presuppositions that are commonly, though perhaps not always consciously, brought to bear by the user. For example, the word "model" in the second title is understood against a substantive background of information about animal models of human disease, and in that context implies that magnesium deficiency causes a disorder resembling epilepsy in the rat. Several hundred analogous title pairs were examined in the course of the migraine-magnesium study, for most of which the linkage was less obvious than in the case above. The user often must make just an educated guess as to which leads are most promising (Swanson, 1991).

The problem we identify in this example therefore is not how or whether to draw an inference about the possible effect of magnesium on migraine, given the above two titles, but rather how these two titles (or Medline records), and other pairs analogous to them, could have been found and brought together in the first place without knowing in advance about any specific link such as epilepsy. That task cannot be done using only a conventional Medline search. However, if one first uses Medline to form a local file consisting of all titles with "migraine," and a second file that consists of all titles with "magnesium," then a

straightforward computer procedure can produce a list of all words common to the two sets of titles. "Epilepsy" would be on the list. One can think of this procedure, which Arrowsmith takes as its point of departure, as a "higher order Medline search." Arrowsmith then automatically filters out noninteresting words (by means of an exclusion list, or stoplist, compiled in advance and built into the system), makes certain morphological transformations (such as plural to singular), constructs and matches phrases, and otherwise exploits information from the Medline record to juxtapose pairs of text passages for the user to consider as possibly complementary. (Arrowsmith can process abstracts as well as titles but, for files of more than 1,000 or so records, it is more efficient and more effective to search, download, and subsequently examine just titles. The restricted context makes it easy to see and assess the A-B relationships when both A and B are in a title and similarly for B-C.) Any inferences about the significance or nature of the linkage between the above two titles, once they have been brought together, are left to the user. Arrowsmith, by creating suggestive juxtapositions of database records, is an aid to scientific discovery but not in itself a mechanism of scientific discovery.

Automatic Generation of a Candidate List for A

Arrowsmith can also do more than help uncover linkages between an initially given A and C. Assume that at the outset only C, the disease under investigation, is given, and the user does not have in mind a specific hypothesis for A (an agent that might act as cause or cure). Then, instead of a specific A, a broad category (AA) may be chosen; such a choice can be simple and effective. In general, categories of exogenous substances that may enter the body and might conceivably have beneficial or adverse effects on C are of interest. Especially important are dietary factors (or deficiencies), toxins, and categories of pharmaceutical agents or their targets (Swanson, 1991). Arrowsmith can then begin with Medline files for C and AA and from these derive a list of specific candidates for A. For example, Arrowsmith was able to start with pre-1988 literature on "migraine" as C, use a category based on dietary or deficiency factors (AA), and produce "magnesium" as a top-ranking candidate for A (Swanson, 1991; Swanson & Smalheiser, 1997).

Direct A-C Search as First Step

It is important for the user who wishes to investigate indirect or implicit connections between A and C to understand that the first step—prior to using Arrowsmith—is to find all articles that are explicitly about both A AND C by means of a conventional or "direct" Medline search. Insofar as indirect linkages are already known (i.e., published), one would expect to find a discussion of them in articles belonging to the A-C

intersection. Failure to understand the contents of the A-C intersection may result in failure to distinguish new from old in the Arrowsmith output.

To conduct a good direct search, some skill and experience with Medline searching is required and in particular familiarity with the medical subject heading (MeSH) hierarchical structure, the superimposed subheading structure, and the organization of the Medline record. Searching of other major biomedical databases, including BIOSIS, EMBASE, and the Science Citation Index, is also important. In some cases, the existence of a sizable direct literature does not necessarily imply that the A and C literatures are well-integrated. For example, in our study of magnesium in the central nervous system, we found a substantial direct literature. But a citation analysis revealed a highly fragmented structure, not at all characteristic of researchers investigating a common problem who cite each other, and are co-cited, extensively (Smalheiser & Swanson, 1994, pp. 5-8). In other cases, we encountered small direct literatures that have never been cited at all in one or the other of the A or C literatures, indicating that new connections were published but then ignored. Our experience underscores the importance of conventional database searching and citation analysis prior to using Arrowsmith for a literature-synthesis study.

In any event, in the more straightforward case in which a well-constructed direct search turns up little or nothing in any of the major appropriate databases, a conventional database search cannot then go any further toward discovering unknown indirect links such as epilepsy in the above example. Arrowsmith is designed to solve that problem. We next explain what Arrowsmith does and how to use it on the Internet.

Arrowsmith on the Web

Arrowsmith may be used free of charge at the Web site: http://kiwi.uchicago.edu. The input to Arrowsmith consists of two files that the user first creates by searching Medline and downloading the resulting records to the user's local computer. We refer to these two local files as File A and File C, both of which must then be transmitted to the server kiwi.uchicago in order to be processed by Arrowsmith. Uploading local files to a remote server can be implemented using Netscape.

Preparing the Input Files

The user begins with some problem (which may be a medical disorder of unknown cause, such as migraine) and conducts a Medline search for records about that disorder (a title-word search is preferable for large files), then downloads the resulting records or titles to a local File C. Similarly, a second Medline search creates a target literature, A (such as magnesium), or some broader category (AA), that is downloaded to File A.

The intersection A AND C is presumed to have been investigated beforehand as noted above. The Arrowsmith software operates in five stages. The user normally will exit after each stage and reconnect at a later stage when results are ready (e-mail addresses are used to identify individual files and results).

Stage 1: Transmitting the Two Input Files to the Server kiwi.uchicago

The kiwi Web site is designed to accept large files transmitted by Netscape. The user provides the local pathname/filename. After Arrowsmith receives File C and File A, it creates a list of all "important" words and phrases common to the two files. This list of terms provides the source for intermediate linkages (B) between A and C. The distinction between words that are "important" and words that are not is implemented by means of a large stoplist (words to be excluded) compiled in advance by applying human judgment and then built into Arrowsmith for all applications. Certain variant word forms are also matched. The output of this stage is a preliminary list of B-terms made available to the user (at Stage 2) five to thirty minutes (depending on file sizes) after Files C and A are received.

Stage 2: Editing the B-List

The preliminary B-list may contain several hundred terms and should be edited by the user. Notwithstanding the stoplist filter, the B-list often contains many terms that the user would not consider of potential interest as linkages in light of the particular problem at hand. At the Web site, the preliminary B-list appears in a scrollable "option" window that permits multiple selection of terms. The selected (highlighted) terms are then automatically deleted from the B-list.

Stage 3: Organized Display of Medline Records as Output

The edited B-list is displayed in a window in which each B-term is a pointer to the subset of Medline records from File A containing that term, a subset called the "AB" records. Each AB display contains a pointer to the corresponding set of BC records, thus facilitating a systematic, organized process of point-and-click browsing of Medline records. For each B-term, the corresponding AB records are, in effect, juxtaposed with BC records to help the user notice a possible A-C relationship. Successful use of Arrowsmith depends on the user's subject knowledge, ingenuity, and ability to see promising connections suggested by comparing AB records with BC records for each B, as illustrated earlier in comparing a magnesium-epilepsy title with an epilepsy-migraine title.

An online example of Arrowsmith title-browsing has been prepared as an interactive demonstration (dem2) at Stage 3 of the kiwi Web site. The example is based on 2,800 migraine titles and 8,000 magnesium titles (all pre-1988, the time frame of the original study [Swanson, 1988]). The computer-produced B-list consisted of 260 terms and was edited manually

to about 100 terms. The user may click on any term in the B-list to see the corresponding magnesium titles, then click on BC to see the migraine titles for that same B-term. The next two stages show what can be done if the user had not considered magnesium at the outset as a possible solution to the migraine problem.

Stage 4: Ranking Individual "A" Terms

Stages 4 and 5 do not apply if File A, above, was based on a specific substance (such as magnesium). However, if File A was created by searching a broader category (such as dietary substances), then we refer to it here as File AA and it becomes of interest to identify more specific A-terms that occur in the records within the AA category. Arrowsmith derives, from the AAB records, a list of words and phrases that become candidates for these more specific terms. The list of candidates is called the A-list. Each term on the A-list is associated with all B-terms that co-occur with it in the AAB records.

The A-list terms are then ranked by the number of their associated terms from the B-list. This method is a simplified version of the ranking method discussed in Swanson and Smalheiser (1997). Thus, the output of Stage 4 is a (preliminary) ranked A-list. Returning to our example using "migraine" to create File C and a dietary/deficiency category to create File AA, the word "magnesium" appeared at the top of the resulting A-list.

Stage 5: Editing and Grouping Terms on the A-List

As was the case for the B-list, the A-list may contain many terms of no interest that should be manually deleted, and it may contain synonyms or related terms that should be grouped together for purposes of ranking. Stage 5 presents the A-list within a scrollable option window that permits multiple selection. Two modes of operation are offered—a deletion mode and a grouping mode. In the first mode, all terms selected are deleted just as in Stage 2. In the second mode, all terms selected by the user are grouped together and treated as synonymous for the purpose of ranking. For example, the A-list might contain ascorbate, ascorbic acid, and vitamin C. In one pass through the window, clicking on these three terms will create a group in which all associated B-terms from each of the three are combined into a single new total; repeating the ranking procedure then gives the group a higher rank than any of its component A-terms. Or the user may choose to form a broader grouping such as all terms that refer to antioxidants, which would include the vitamin C terms above. Alternation between the deletion mode and the grouping mode is permitted using each mode as many times as desired. The final A-list is then reranked.

Nothing in the foregoing process determines whether any term on the A-list does or does not co-occur directly with C in Medline records;

such co-occurrence should be separately determined by means of a conventional Medline search. Extensive co-occurrence probably indicates that the relationship with C is already well known, and so the A-term in question may not be of further interest (however, see the earlier discussion of the direct search and the possibility of encountering fragmented structures).

The sole purpose of the A-list is to offer some automatically generated promising choices of specific A-terms for the user's consideration. Once the user has chosen a single specific A (such as magnesium) that seems promising as the basis for File A, then the next step is to re-run Arrowsmith beginning again at Stage 1. The category restriction may be omitted altogether (thus leading to the largest B-list for the A and C under consideration) or it (or perhaps a revised version) may be included as part of the Medline search that creates File A.

Synonym Recognition and the Role of Medical Subject Headings (MeSH)

The heart of Arrowsmith is the computerized process of finding and matching words and phrases that occur in both input files (Files A, C) as an approach to helping the user identify complementary passages of text from titles or abstracts. In addition to matching identical terms, Arrowsmith also matches certain morphological variants, including most cases of singular versus plural, and it can identify synonyms insofar as they are indexed by a common subject heading (MeSH). To take advantage of the synonym matching capability, MeSH terms must be included for each record in the input files A and C.

The output of the matching process consists of a list of terms (the B-list) that itself may contain synonyms or context-dependent equivalencies that the user may wish to take into account. A future version of Arrowsmith will provide more assistance by presenting to the user a list of word (and phrase) pairs that are candidates for synonyms or "surrogate synonyms" (sometimes called "searchonyms") that could serve as an aid to editing (Stage 2, 5), browsing (Stage 3), and forming groups (Stage 5). Words will be paired if they tend to appear in similar contexts as defined with the help of statistics based on second order co-occurrence data. Two words that are synonymous or equivalent tend not to co-occur in a highly restricted context such as a title and so do not have a strong first order title co-occurrence correlation. But their tendency to occur in similar contexts gives rise to relatively stronger second order title co-occurrence correlation.

Synonyms, searchonyms, variant word forms, and co-occurrence statistics can at best provide only a partial solution to the difficult problems of detecting complementary or suggestive pairs of text passages, but Arrowsmith is especially valuable for developing and testing improved approaches and techniques.

Patterns of Complementarity and Suggestivity

"A causes B, B causes C; hence A causes C" can be taken as a paradigm for complementarity, but it is an idealization. As we have gained experience using Arrowsmith, it has become clear that transitivity is almost never assured, and we have to settle for the less formal and less tidy idea of suggestibility (Swanson, 1991). The problems of suggestivity and complementarity as expressed in natural language text are complex and subtle. Nonetheless, Arrowsmith is now able to produce large numbers of suggestive juxtapositions of Medline titles or records, and it is reasonable to expect further improvement with the accumulation of additional inelegant ad hoc empirical rules with little else to recommend them except that they seem to work.

In studying links that actually occur in the natural language text of title words and phrases, we have identified a few regularities or patterns that may become the basis for useful rules. For example, the A-B and B-C relationships largely fall into three groups that can be called "influence," "similarity," and "focus." The concept of "influence" (of A on B or B on C) can be expressed by many different words, including: increases, decreases, attenuates, reduces, promotes, inhibits, ameliorates, exacerbates, enhances, causes, accelerates, facilitates, triggers, catalyzes, competes with, interferes with, or acts synergistically. The direction of influence may also be reversed, with B influencing A. The software is indifferent and symmetric with respect to the direction of any relationship. The concept of "similarity" can be important either alone (A is similar to B and B is similar to C) or in conjunction with "influence": A influences B and B is similar to C, thus suggesting that A might influence C (e.g., magnesium deficiency triggers or exacerbates epileptiform seizures; migraine in some respects is similar to epilepsy, suggesting therefore that magnesium deficiency may trigger or exacerbate migraine attacks). The category "focus" refers loosely to a cluster of relationships between some disease and its manifestations in specific cell types, processes, mechanisms, pathways, markers, and organs, or at any anatomic locale at which a focal pathology is a characteristic feature. "A" may be a drug or other substance that is active at such a focus, B, in which case the "influence" category probably applies. The relationship between a disease and its manifestations (e.g., pathologic markers for it) may be more difficult to categorize, so "focus" is used simply as a tentative collective name for possibly several types of relationship.

For example, indomethacin inhibits a variety of cholinergic responses; cholinergic deficits are characteristic of Alzheimer's disease. (Thus indomethacin, which is thought to have a protective effect in Alzheimer patients on the basis of clinical trials, might also have unexpected adverse effects [Smalheiser & Swanson, 1996a].)

The foregoing regularities notwithstanding, natural language is richly

expressive, and the variety of ways in which meaningful biological linkages can be suggested to the expert human observer may be so great as to defeat any attempt to formalize and automate the recognition and inference process. Arrowsmith in its present form does not attempt to do so but instead is designed to organize and display records so as to facilitate human recognition of implicit connections. Investigating patterns of complementarity, however, may lead to richer and improved displays of information to the user and so perhaps to improved stimulation of hypotheses. Arrowsmith is not only a practical tool that can aid the biomedical researcher, it is also a research tool for investigating the problems of finding and identifying natural language text linkages.

The Role of Human Intelligence

At several points in the procedure, Arrowsmith receives a boost from human input that helps it perform as if it were intelligent. The first boost is the choice of the problem and its literature C, plus the choice of A as a specific target, or AA as a more general target category. Using A, AA, and C to construct a good Medline search also requires knowledge, experience, and judgment at the outset. The second boost is the stoplist filter, which greatly reduces the number of useless connections that otherwise would clutter the output. The stoplist is compiled using human judgment (and guesswork) concerning which words probably could not play any useful role in forming biologically meaningful and helpful linkages. It is intended as a one-time compilation, not ad hoc for each Arrowsmith application, but the stoplist does grow as the human compiler gains experience with Arrowsmith and now includes about 7,000 words. The remaining boosts come from the user in editing the B-list and A-list and in forming groups within the A-list. Finally, given the juxtaposed AB-BC titles or abstracts, any identification of promising implicit linkages of biological importance depends on the knowledge and perspicacity of the user.

Assessments of Arrowsmith by Others

This project has been analyzed, enhanced, and extended in a number of recent papers (Chen, 1993; Cory, 1998; Davies, 1989; Finn, 1998; Garfield, 1994; Gordon & Lindsay, 1996; Gordon & Dumais, 1998; Kostoff, 1998; Rikken, 1998; Spasser, 1997). Analogous work on computer-generated discovery in chemical reaction pathways has also been reported (Valdes-Perez 1994). Valdes-Perez (1999) has assessed four successful computer-assisted discovery programs in chemistry (MECHEM), medicine (ARROWSMITH), mathematics (GRAFFITI), and linguistics (MPD/KINSHIP). He explains why he believes that each of them has produced results that are novel, interesting, plausible, and intelligible.

REFERENCES

Chen, Z. (1993). Let documents talk to each other: A computer model for connection of short documents. *Journal of Documentation, 49*(1), 44-54.

Chevalley, T. H.; Rizzoli, R.; Manen, D.; Caverzasio, J.; Bonjour, J. -P. (1998). Arginine increases insulin-like growth factor-I production and collagen synthesis in osteoblast-like cells. *Bone, 23*(2), 103-109.

Corpas, E.; Harman, S. M.; Blackman, M. R. (1993). Human growth hormone and human aging. *Endocrine Reviews, 14*(1), 20-39.

Cory, K. A. (1997). Discovering hidden analogies in an online humanities database. *Computers and the Humanities, 31*(1), 1-12.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation, 45*(4), 273-301.

Finn, R. (1998). Program uncovers hidden connections in the literature. *Scientist, 12*(10), 12-13.

Garfield, E. (1994). Linking literatures: An intriguing use of the *Citation Index*. *Current Contents, 21*, 3-5.

Gordon, M. D., & Lindsay, R. K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science, 47*(2), 116-128.

Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science, 49*(8), 674-685.

Hurson, M.; Regan, M. C.; Kirk, S. J.; Wasserkrug, B. A.; & Barbul, A. (1995). Metabolic effects of arginine in a healthy elderly population. *Journal of Parenteral and Enteral Nutrition, 19*(3), 227-230.

Kirk, S. J.; Hurson, M.; Regan, M. C.; Holt, D. R.; Wasserkrug, H. L.; & Barbul, A. (1993). Arginine stimulates wound healing and immune function in elderly human beings. *Surgery, 114(2)*, 155-160.

Kostoff, R. N. (1998). *Science and technology innovation.* Retrieved January 1, 1999 from the World Wide Web: http://www.dtic.mil/dtic/kostoff/Swanson2.txt.

Rikken, F. (1998). *Adverse drug reactions in a different context: A scientometric approach towards adverse drug reactions as a trigger for the development of new drugs.* Unpublished doctoral dissertation, Rijksuniversiteit Groningen.

Ross, B. M. (1998). In reply. *Archives of General Psychiatry, 55*(8), 753.

Spasser, M. A. (1997). The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science, 48*(8), 707-717.

Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience Research Communications, 15*(1), 1-9.

Smalheiser, N. R., & Swanson, D. R. (1996a). Indomethacin and Alzheimer's Disease. *Neurology, 46*(2), 583.

Smalheiser, N. R., & Swanson, D. R. (1996b). Linking Estrogen to Alzheimer's Disease: An informatics approach. *Neurology, 47*(3), 809-810.

Smalheiser, N. R., & Swanson, D. R. (1998a). Calcium-independent phospholipase A-sub-2 and schizophrenia. *Archives of General Psychiatry, 55*(8), 752-753.

Smalheiser, N. R., & Swanson, D. R. (1998b). Using Arrowsmith: A computer-assisted approach to forming and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine, 57*(3), 149-153.

Swanson, D. R. (1986a). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1), 7-18.

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56*(2),103-118.

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science, 38*(4), 228-233.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine, 31*(4), 526-557.

Swanson, D. R. (1989a). Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy. *Journal of the American Society for Information Science, 40*(5), 356-358.

Swanson, D. R. (1989b). A second example of mutually isolated medical literatures related by implicit unnoticed connections. *Journal of the American Society for Information Science, 40*(6), 432-435.

Swanson, D. R. (1990a). The absence of co-citation as a clue to undiscovered causal connections. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 129-137). Newbury Park, CA: Sage.

Swanson, D. R. (1990b). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association, 78*(1), 29-37.

Swanson, D. R. (1990c). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine, 33*(2), 157-186.

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein (Ed.), *SIGIR '91* (Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois, USA, October 13-16, 1991) (pp. 280-289). New York: Association for Computing Machinery.

Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge, *Library Trends 41*(4), 606-631.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*(2), 183-203.

Valdes-Perez, R. E. (1994). Conjecturing hidden entities by means of simplicity and conservation laws: Machine discovery in chemistry. *Artificial Intelligence, 65*(2), 247-280.

Valdes-Perez, R. E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence, 107*(2), 335-346.

Discovering Hidden Analogies in an Online Humanities Database*

KENNETH A. CORY

ABSTRACT

VOLUMINOUS DATABASES CONTAIN HIDDEN KNOWLEDGE—i.e., literatures that are logically but not bibliographically linked. Unlinked literatures containing academically interesting commonalities cannot be retrieved via normal searching methods. Extracting hidden knowledge from humanities databases is especially problematic because the literature, written in "everyday" rather than technical language, lacks the precision required for efficient retrieval, and because humanities scholars seek new analogies rather than causes. Drawing upon an efficacious method for discovering previously unknown causes of medical syndromes and searching in the Humanities Index, a periodical index included in WILS, the Wilson Database, an illuminating new humanities analogy was found by constructing a search statement in which proper names were coupled with associated concepts.

PROBLEM STATEMENT

Journal articles are logically linked if they deal with the same subject(s). If, within online databases, their citations share common terms, they are also bibliographically linked. Retrieving them is simply a matter of inputting a common term. Unfortunately, many logically connected citations lack common terms. Conventional searching methods cannot retrieve these "noninteractive" citations. Swanson (1986b, 1988), working in medical databases, developed a novel searching technique that retrieved hidden

*This article first appeared in *Computers in the Humanities, 31*, pp. 1-12, 1997 and is reprinted by kind permission of Kluwer Academic Publishers.

Kenneth A. Cory, Wayne State University, Library and Information Science Program, 106 Kresge Library, Detroit, MI 48202
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 60-71

knowledge—i.e., logically connected but bibliographically unconnected citations. The intent of this study was to determine if humanists could apply Swanson's methodology.

Extracting hidden knowledge from humanities databases is problematic because the literature, written in "everyday" rather than technical language, lacks the precision required for efficient retrieval, and because humanists more often seek new analogies rather than causes. This researcher overcame those obstacles by constructing a search statement in which proper names were coupled with associated concepts. The discovery of a previously unnoticed analogy between the epistemological ideas of Robert Frost and the ancient Greek philosopher Carneades suggests that the voluminous contents of online databases may collectively be a new kind of primary source.

Discovering new humanities knowledge is crucially important because humanities scholars rarely have access to new information sources. Natural and social scientists create new knowledge via experimentation. The latter also rely on surveys. And from economic, political, and social statistics, they have replenishing sources on which to base novel conclusions. Humanists rely on primary sources, something not available in humanities databases. Therefore, if humanists could "create" new knowledge by finding links between and among existing citations, that would accelerate their research efforts.

Accelerating the research process is also a practical necessity. Online searching is expensive. Telecommunication connect charges consume roughly half the cost of all searches and apply even when nothing of value is found. Researchers would value any technique, however imperfect, that cost-effectively retrieves worthwhile citations.

Scholars of different humanities disciplines consult different databases, each of which has its idiosyncrasies. To ensure that database peculiarities would not confound the research findings, this study was focused on finding new knowledge in a single electronic literary database.

Theoretical Framework

Summarizing arguments from philosophy and information science, Davies (1989) concluded that the sum of the world's knowledge is vastly more than the sum of all knowledge within publications. Potentially, each concept can generate logical consequences, the results of which cannot be anticipated. This implies the existence of hidden knowledge within databases.

Swanson (1986a, 1986b, 1988, 1989a, 1989b, 1990a, 1990b, 1991, 1993) published a remarkable series of papers calling attention to this reality by finding previously unknown causes of medical syndromes. His success and his challenge to librarians to develop searching methods to find other logically linked noninteractive documents inspired this study. Obviously,

if electronic medical databases contain vast quantities of undiscovered knowledge, so may all large electronic databases, including those in the humanities.

Two Inherent Problems

In planning this study, considerable doubt existed that Swanson's procedure would yield new humanities knowledge. One major problem involves the nature of words appearing in bibliographic citations. In Swanson's methodology, searchers must retrieve a large number of titles in order to identify frequently reappearing words. These words must be specific and descriptive. Because medical terminology is technical and technical terms tend to lack synonyms, medical titles tend to be descriptive of an article's contents. This is a convention of all technical literature. Humanities titles often are nondescriptive, even imaginative. There are few technical terms, and synonyms abound for almost every word. Benaud and Bordeinanu (1995) describe the problem succinctly:

> Several factors make database searching in the humanities particularly difficult. Chief among these is the semantic ambiguity attached to many humanistic terms. The high occurrence of natural language in humanistic writing that impedes the selection of index terms also presents difficulties for bibliographic retrieval. For example, the title *The Mirror and the Lamp*, written by M. H. Abrams in 1953, would not alert the database searcher that he found a work on romantic theory and the critical tradition. (pp. 42-43)

A second problem inheres in the differing nature of the knowledge sought. Medical researchers commonly look for causes. In contrast, humanistic researchers, especially those seeking literary knowledge, commonly seek to provide new interpretations. Stone (1982), explaining the tendency of humanists to work alone, emphasized that "the individual scholar's interpretation is paramount" (p. 294). The subjective nature of an interpretation renders empirical verification moot. Instead, as reported by Wiberley (1991), peer acceptance is the normal criterion of an interpretation. One could say that peer confirmation is the equivalent of hypothesis confirmation in the sciences. For these reasons, information scientists hoping to discover significant new humanities knowledge are seeking intrinsically elusive material. They can do no more than call attention to information appearing to support new interpretations.

Focusing on Hidden Analogies

Because analogies often establish illuminating interpretations, seeking new interpretations commonly requires seeking analogous materials. Analogies are the "comparison of two things, alike in certain respects; particularly a method of exposition by which one unfamiliar object or idea is explained by comparing it to something more familiar" (Holman

& Harmon, 1992, p. 20). The practical value of finding a hidden analogy between certain authors is that knowing the ideas of one may help explain similar ideas of the other.

As humanists know, the weakness of an analogy is that few different objects or ideas are essentially the same to more than a superficial observer or thinker (Holman & Harmon, 1992). Nevertheless, though often meaningless, analogies occasionally form the basis of new interpretations.

Considerable searching failed to locate any systematic attempts to discover hidden analogies. Davies (1989), in his delineation of the categories of hidden knowledge, mentioned that Farradane, as early as 1961, hoped that relational indexing might be developed that would be capable of "recognizing analogies between subjects." Davies (1989) provided a reason to believe that hidden analogies could be found. "According to von Bertalanffy, there are many instances where identical principles were discovered several times because the workers in one field were unaware that the theoretical structure required was already well developed in some other field" (p. 284). Davies's conviction that hidden analogies could be discovered encouraged this investigator to believe that shifting the focus from causation to analogies could uncover hidden knowledge in humanities databases.

Limitations

There are limitations to using analogies:

1. Analogies are subjective concepts. They cannot be laboratory tested. Therefore, any report of new knowledge must be understood as tentative—something appearing to be significant and worth pursuing by humanist scholars.
2. Replication may produce nothing of value. A method is verified if identical results are consistently produced from identical testing conditions. Replicating the method employed in the following search procedure may produce an analogous relationship, but that relationship may not be meaningful.
3. Database searching is an inexact science. Success is partially dependent on the educational level and intelligent imagination of the searcher.
4. Searching for analogies may only be successful when the search is phrased in terms of "Who or what influenced someone or something?" Influence questions are essentially causal in nature. And, as Swanson has demonstrated, unknown causes can be discovered.

Research Methodology

Because of Swanson's success, and because the method he employed to search for hidden transitive relations is also recommended by Davies

(1989) for finding hidden analogies, this researcher proceeded from that method while searching for hidden analogous knowledge in the humanities. This systematic trial-and-error strategy was best described by Davies (1989):

1. A search statement is constructed based on the subject under investigation.
2. A lengthy list of title citations is retrieved.
3. Titles are examined for recurring words or phrases. These words or phrases must not be synonymous with the original subject.
4. These recurring terms are used, one at a time, to construct a search statement for a second round of searching. In the second round, the original search term is omitted from the search statement.
5. The resulting titles are examined. Here, the researcher makes a strategic guess. Articles with titles that seem logically related to the original subject are retrieved and read. The researcher attempts to determine whether or not the contents of particular articles might significantly illuminate the original subject. If so,
6. The researcher attempts to determine whether or not an article is bibliographically linked to the original subject. As described below, this is done by conventional searching (p. 294).

A Specific Example of the Swanson Method Applied to a Medical Database

Swanson identified magnesium deficiency as a causal agent in the occurrence of migraine headaches. At that time, the causes of migraine headaches were unknown.

His original search term was "migraine." That produced a plethora of title citations containing "migraine." These were examined for recurring words or phrases. Among these were "calcium entry blockers" and "platelets." They were included in a second round of searching in which migraine was omitted. Eventually, searching produced several sets of citations, with two citations in each set. In each of the first list of titles, "migraine" appeared, along with a chemical or condition known to be present in migraine sufferers. In each of the second list of titles of the corresponding set, the same chemical or condition appeared but not migraine. In the following examples, note that magnesium is not mentioned in relation to migraines but does appear in reference to a condition known to be associated with migraine:

a: Role of *calcium entry blockers* in the prophylaxis of migraine
b: Magnesium: nature's physiologic *calcium blocker*

a: Evidence of enhanced *platelet aggression* in platelet sensitivity in migraine patients
b: Protective effects of dietary calcium and magnesium on *platelet* function

Swanson (1993) concluded: "Because of the shared 'linkage' terms shown in italics, each of the. . . pairs of titles raises the question of whether magnesium deficiency might be implicated in migraine." He labeled these sets "'complementary literatures,' [or] a pair in which one literature appears to contain a potential solution to a problem posed in the other" (p. 620). Bringing together complementary citations allows even a nonexpert to notice a possible causal relationship. Eventually, one must read the articles. Merely matching citations is insufficient.

Application of Swanson's Methodology to Humanities Databases

Overcoming the Limitations of Humanities Language

Because of the "ordinary language" found in humanities citations, a searching method had to be developed using substitutes for recurring words and phrases. The one type of recurring term found in humanities citations for which there are virtually no synonyms is names.

Using Names to Control Imprecise Searches

Names meet the crucial criteria of Swanson's searching method: they recur and, because they commonly refer to a single person, they are specific. Names are not ambiguous. Names have no synonyms. Of course, a name may be logically linked to more than a single concept. Searching with "wagner, richard" might retrieve citations pertaining to the development of the leitmotif, Tannhauser, or revolutionary activity in nineteenth-century Germany. Admittedly, names do lack the precision of medical terms. Nevertheless, using names to construct search statements considerably reduces the ambiguity inherent in humanities terminology. Tibbo (1991) quotes Wiberley who, "in a study of terms taken from encyclopedias and dictionaries in the humanities, confirms the importance of singular proper terms, especially the names of persons. He concludes that subject access is far more straightforward than has been recognized if subjects are expressed through such proper names" (p. 300).

Because names are often associated with multiple concepts, they cannot, by themselves, be used as search statements when attempting to discover new knowledge. However, when names are associated with specific concepts, they can serve as controlling terms that direct a second round of searching.

The following is a detailed example of how three graduate students discovered a new analogy. Note a major departure from Swanson's method. These researchers did not look for recurring words or phrases. Any name, even if mentioned only once, may be profitably used in the second round of searching. The important precepts are to: (1) couple the name with an associated concept, and (2) omit the original term in the second searching round.

Carneades/William James/Robert Frost

Methodology

Responding to an inquiry about Robert Frost (1874-1963), student researchers produced several citations pertaining to the nineteenth-century American pragmatic philosopher, William James (1842-1910). Subsequent reading revealed that Robert Frost had definite philosophical convictions pertaining to how truths can be verified. His ideas were influenced by James, so much so that knowing the underlying philosophy of James clarifies the epistemological ideas of Frost.

An intriguing question evolved: Could there be an unknown literary or philosophical antecedent of James, the discovery of whom might assist in understanding Frost? Certainly humanists are aware that the works of Frost are a compendium of the works of earlier thinkers in addition to James. Even those who may not have shaped Frost's thinking directly may have had a significant influence and would thus be worth knowing about because they can assist one to understand Frost's ideas. The question became: What author(s) not known to have directly influenced Frost's poetry have nonetheless indirectly influenced it via someone else? Would knowing the ideas of this author contribute to a better understanding of Frost?

Frost became the *C* in the equation: Unknown author *(A)* → James *(B)* → Frost *(C)*. The algebraic equation, If $A = B$, and $B = C$, then $A = C$, is true for numbers. For humanities connections, it is sometimes meaningful but usually not. Nevertheless, it is a worthwhile mental model for seeking logical connections among citations.

Results from preliminary searching indicated that the most promising database was the Humanities Index, one of eight periodical indexes included in the Wilson database, also known as WILS. The Modern Language Association's MLA-CD was rejected as it lacks pure philosophy, which limits its usefulness for searching about William James. The Humanities Index was selected because of the breadth of its contents: folklore, history, language and literature, literary and political criticism, performing arts, philosophy, and religion and theology. Coverage is extensive with articles from over 1,000 periodicals indexed. Moreover, its time coverage, 1983 to the present, is fairly lengthy and is updated monthly.

In the second round of searching, the name "William James" was employed. The intention was to retrieve names associated with James that would suggest promising avenues of additional searching. As stated above, the retrieved names must be associated not only with the subject under investigation—i.e., James—they must also be coupled with a concept common to Frost and James. After reading that Frost's interest in James was epistemological—i.e., he accepted James's methods of verifying ideas—epistemology was chosen as the associated concept. The search statement was "james and epistemology." That produced only thirty useful citations. The search needed to be broadened. Clearly, "epistemology"

by itself was insufficient. A quick inquiry into a standard reference source was informative. James's epistemological thinking led him to pragmatism. Therefore, pragmatism was included in the next search (Duran, 1953, pp. 510-13). The search statement "james and (epistemology or pragmatism)" produced eighty citations.

Retrieved titles not containing a name associated with James were eliminated. Forty-four citations containing names associated with James and pragmatism looked promising. These included: Carneades (circa 214-129 BCE), Bernard of Clairvaux (eleventh century), John Dewey (1859-1952), and C. S. Pierce (1839-1914). Conventional searching—i.e., a keyword search: "k = [name of an author appearing in a retrieved citation] and frost" demonstrated that there were bibliographic links between Frost and the other authors except Carneades.

Two citations seemed promising. One indicated a logical link between the pragmatic philosophy of Carneades and James and the other indicated a logical link between the pragmatic philosophy of James and Frost. These formed a set:

- Doty, R. (1986). "Carneades, a forerunner of William James's pragmatism." *Journal of the History of Ideas, 47*(1), 133-138.
- Shaw, W. D. (1986). "The poetics of pragmatism: Robert Frost and William James." *The New England Quarterly, 59*(2), 159-188.

Note that "James" and "pragmatism" appear in each citation. Also, note that "Carneades" appears as a forerunner of James and James appears as a forerunner of Frost. All three names are logically linked to pragmatism with William James being common to both. Reading the articles verified that both dealt with epistemological subjects. These citations are bibliographically unlinked as indicated by the fact that nothing in the Doty article refers to Frost; nothing in the Shaw article refers to Carneades. Therefore, conventional searching could not retrieve these citations together. That is, searching by "frost, robert" would retrieve citations pertaining to James but not to Carneades.

Next, the searchers had to determine if the logical connection was already known. They determined that it was not, by searching to see if either of the authors of the set cited each other (Doty citing Shaw or Shaw citing Doty in an article(s) about Carneades and Frost). Also, they checked to see if any third author cited both of them (x cites both Doty and Shaw in an article about either Frost or Carneades). Because both searches produced negative hits, the connection was presumed to be unknown among academicians.

Analysis of the Articles

Briefly analyzing the retrieved articles helps one appreciate the necessity of subjective insight. Although there is an efficacious methodology

to report, its efficacy depends upon a well-educated searcher. In the humanities, a single name, identified as relevant, can uncover a significant analogy provided that the searcher can perceive a logical link between citations. A searcher need not be a refined scholar, but he or she must be academically knowledgeable.

Carneades, James, and Frost

Doty (1986) contends that Carneades and James developed comparable theories of truth. Doty found no evidence proving that James had read Carneades. He accounts for their parallelism by proposing that both men faced similar opponents and reacted in like fashion. Carneades founded the New Academy, which espoused skepticism as an alternative to stoicism. Similarly, James, reacting against rationalist philosophy, developed his theory of pragmatism.

Carneades believed that individuals do not perceive certainty or "truth" in their experiences—or at least what they believe to be truth. In order to determine the validity of experience, Carneades developed three criteria: the probable, the irreversible, and the tested. These criteria parallel James's correspondence test of truth, coherence test, and pragmatic truth-test. Thus, both men "present a truth-test consisting of the verification of a hypothesis by empirical methods" (p. 136). However, James's test is one of truth presumed to be knowable, and truth for him is the product of empirical verification. Carneades, on the other hand, rejects truth as being beyond human knowing and offers a test of probability. Despite that difference, Doty's article clearly establishes a logical link between Carneades and James.

Next, reading Shaw's (1986) essay described how James influenced Frost. Frost had extensively read James. Frost's poems often explored the consequences of James's pragmatic concept of truth. To Frost, the "possession of true thoughts means everywhere the possession of invaluable instruments of action" (quoted in Shaw, 1986, p. 161). Shaw contends that one of the most prominent features of this definition is the impossibility of developing a philosophically or critically interesting theory concerning the dictionary or essential meaning of a word. This establishes an apparent congruence of thought between Carneades and Frost. Further reading revealed that, throughout the poems of debate, Frost has one of his characters substitute a blueprint for action for a conventional dictionary definition. This corresponds to a statement of James's about being lost in the woods and how the true thought of a house is useful "because the house which is its object is useful" (quoted in Shaw, 1986, p. 162).

James did not believe that a pragmatic definition required an experiment to prove that it was true, and Frost exemplifies this concept in a number of poems such as "The Mountain." In this poem, Frost has his farmer establish conditions under which his statement could be verified—yet he does not then actually have the farmer verify the statement.

James often used the term "tough-minded" to describe the skeptical and empirical temperament that he admired. In his poems, Frost shows "tough-minded" speakers (i.e., skeptics) debating with what he called "tender-minded" speakers who were dogmatic idealists. Frost parallels James in his belief that "all attempts to ground practice in traditional theology or metaphysics are attempts to make a god-term of some useless wheel that plays no active part in the cosmic mechanism" (Shaw, 1986, p. 175). Frost exhibits this belief in several poems, most notably in "A Masque of Mercy" and "A Masque of Reason." He argues instead for a workable theism that possesses three qualities in which he believed. Theism, Frost proposed, must be open and free, must be plural, and must be purposive and partly intelligible.

Frost and James also agree upon the notion of freedom. Both believe that freedom exists only when one has to make choices that will produce results that are mutually exclusive. This is evident in "The Road Not Taken." In what are collectively called his poems of departure, Frost also shows the benefits and problems of having the freedom to choose.

The searchers concluded by postulating a logical connection between Carneades and Frost. They reported that Carneades's three criteria are similar to James's three tests of truth. Furthermore, Frost utilized those tests in his poetry. For example, the farmer in "The Mountain" employs two of Carneades's criteria when he sees a stream (the probable) and discusses the possibility of it originating at the top of the mountain (the irreversible). Thus it seems possible to analyze Frost's poetry by employing Carneades's method of verification.

The student researchers were academically reluctant to insist that their finding qualified as an important new analogy. They did claim, and justifiably in the opinion of their professor, sufficient grounds for recommending that humanist scholars read Carneades's works and compare them to the poetry of Frost. The novel idea that the epistemological tests of an obscure ancient Greek may serve as an illuminating philosophical prefigurement of Frost's epistemology does seem worth pursuing. More importantly, whatever the scholarly value of this particular finding, the fact that a previously unrealized analogy has been discovered establishes the efficacy of the described searching method.

Preliminary Conclusion

This study had two objectives: (1) to develop a methodology that would accelerate humanities research by discovering significant hidden analogies within electronic humanities databases; (2) to prove the efficacy of that method by actually discovering a significant analogy that could not be discovered by means of conventional searching.

The result was encouraging. The investigators did link logically related citations that were bibliographically unlinked. "Related" is a subjective

decision, but the principal investigator is confident that sufficient supporting evidence was furnished to make a reasonable case that an important analogy was discovered. And, of course, the primary objective of this project was not to actually discover hidden knowledge but to discover a means of accelerating humanities research via online searching. That has been accomplished.

In any discipline, the possibility of discovering hidden or unlinked knowledge offers improved services, substantial financial savings as compared with trial and error research, and increased status for librarians involved in successful searching. Moreover, for researchers who enjoy both searching challenges and academic subject matter, searching for undiscovered public knowledge offers a new source of personal intellectual excitement.

Future Plans and Recommendations for Future Research

Additional searches are required to confirm and to refine the method in various humanities disciplines. Librarians and/or information scientists are invited to apply Swanson's method or the variation described herein in all subject disciplines, including the natural and social sciences. If, as has been demonstrated in this study, the method can work for humanities topics, it can certainly be employed in disciplines using more technical terminology and in which causation is an important question.

This author is active in the imaging industry. Imaging is the conversion of information from paper, microform, photographic, or voice format into digital format. Imaging vendors are constructing a multitude of large business, scientific, and technical databases that will contain more information than can be retrieved by engaging conventional searching methods. Especially for companies involved in solving technical or environmental problems, information managers searching for new knowledge will find unexpected benefits. Eventually, online databases may be perceived less as static information repositories and more as knowledge generating machines.

Concluding Note

The student investigator, Mark Bowden, offered an appropriate concluding perspective:

> Perhaps humanities scholars will find the greatest benefit of this method is the way it orients one's view of a subject toward aspects or relationships previously unknown. New investigative avenues are opened, new hypotheses are formulated, and new syntheses proposed. At the very least it is a method for scholars to devise original research ideas; at the most it is a powerful tool for revealing hidden connections between persons, places, and events.

ACKNOWLEDGMENTS

This project was funded by a grant from the Wayne State University Humanities Center. The author is indebted to Char Watch for her tireless online searching and to Melissa A. DeNamur, Gretchen Harmor, and Mark A. Bowden, graduate students in Wayne State University's Library and Information Science Program, for their original work and for the work done after the process was reinvented. The author is also grateful to Don R. Swanson for reading a draft of this paper and for his valuable corrections.

REFERENCES

Benaud, C.-L., & Bordeianu, S. (1995). Electronic resources in the humanities. *Reference Services Review, 23*(2), 41-50.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation, 45*(4), 273-301.

Doty, R. (1986). Carneades, a forerunner of William James's pragmatism. *Journal of the History of Ideas, 47*(1), 133-138.

Durant, W. (1953). *The story of philosophy: The lives and opinions of the greater philosophers* (2d ed.). New York: Simon & Schuster.

Holman, H. C., & Harmon, W. (1992). *A handbook to literature* (6th ed.). New York: Macmillan.

Shaw, W. D. (1986). The poetics of pragmatism: Robert Frost and William James. *The New England Quarterly, 59*(2), 159-188.

Stone, S. (1982). Humanities scholars: Information needs and uses. *Journal of Documentation, 38*(4), 292-313.

Swanson, D. R. (1986a). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1), 7-18.

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56*(2), 103-118.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine, 31*(4), 526-557.

Swanson, D. R. (1989a). Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy. *Journal of the American Society for Information Science, 40*(5), 356-358.

Swanson, D. R. (1989b). A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science, 40*(6), 432-435.

Swanson, D. R. (1990a). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association, 78*, 29-37.

Swanson, D. R. (1990b). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine, 33*(2), 157-186.

Swanson, D. R. (1991). Complementary structures in disjoint science literature. In A. Bookstein (Ed.), *SIGIR '91* (Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval: Illinois, USA, October 13-16, 1991) (pp. 280-289). New York: Association for Computing Machinery.

Swanson, D. R. (1993). Intervening in the life cycle of scientific knowledge. *Library Trends, 41*(4), 606-631.

Tibbo, H. R. (1991). Information systems, services, and technology for the humanities. *Annual Review of Information Science and Technology, 26*, 287-346.

Wiberley, S. E. (1991). Habits of humanists: Scholarly behavior and new information technologies. *Library Hi Tech, 9*(1), 17-21.

A Passage Through Science: Crossing Disciplinary Boundaries

HENRY SMALL

ABSTRACT

A METHODOLOGY IS PRESENTED FOR CREATING pathways through the scientific literature following strong co-citation links. A specific path is described starting in economics and ending in astrophysics traversing 331 documents. Special attention is given to where the path crosses disciplinary boundaries and how analogy can be used to model the thought processes involved in such transitions. Implications of information pathways for retrieval, the unity of science, discovery, epistemology, and evaluation are discussed.

INFORMATION RETRIEVAL AND INFORMATION TRANSITIONS

A great deal of information science is concerned with retrieving all the documents from a database that precisely match a user's query. In this magic bullet model of information retrieval, the documents retrieved will ideally be homogeneous in character. Such an ideal is, of course, rarely achieved. In practice, a wide array of documents of varying relevance is retrieved, resembling more an ecology of information than a uniform set.

Less often under consideration is how to understand the diversity and breadth of information that most queries generate, how one topic relates to another, or the transitions from one document to another. Questions such as these naturally arise for large samples of documents and especially multidisciplinary databases. For example, a user interested in a topic such as asthma might retrieve a large number of hits and find that

Henry Small, ISI, 3501Market Street, Philadelphia, PA 19104
LIBRARY TRENDS, Vol. 48, No. 1, pp. 72-108

some deal with treatment options, age factors, psychological aspects, hereditary tendencies, environmental factors, and so on. The question is how to make sense of this diversity.

One reason questions of subject diversity do not come up more often is the tacit assumption that topics or subjects are relatively isolated and distinct from one another, each representing a more or less separate homogeneous entity. Another reason is the assumption that users' information needs are simple and highly specific. This contrasts with the view that information seeking is more like a gradually unfolding discovery process in which the initial query is only the first step in a long journey, each step depending on what came before (Kuhlthau, 1999).

Information Transitions and the Unity of Science

Earlier discussions of the unity of science (Neurath, 1938) or its modern incarnation in E.O. Wilson's (1998) consilience, view scientific knowledge as an interconnected fabric of fields and disciplines. In the sociology of science, it is commonplace to say that a great deal of scientific and technological innovation takes place at the boundaries between disciplines (Lemaine et al., 1976) or by individuals who have crossed from one field to another. Cross-fertilization of fields is another term for this, when an idea in one field finds fertile ground in a neighboring field (Crane, 1972). Information scientists have begun to explore these issues by attempting to find unconnected subject areas which, if connected, might yield new discoveries (Swanson & Smalheiser, 1997). Attempts to visualize information spaces also address subject connections since a visualization must depict the relationships among diverse sets of documents (White & McCain, 1997). It seems likely that future information retrieval systems based on the visual paradigm will have the equivalent of road signs telling the user what direction to travel to reach a particular topic.

Citations and the Structure of Science

One of the best ways of studying the connectedness of information is to use reference or citation links. While connections can also be established by shared vocabulary or indexing terms, a citation link represents a more direct author-selected dependency. By taking a wide-ranging sample of documents across many fields, the unity of scientific information can be examined from a global perspective.

Vannevar Bush's (1945) idea of associative information trails is a natural consequence of the unity of science and the connectedness of knowledge. Hummon and Doreian (1989) attempted to demonstrate this on a small scale by finding a critical path through a DNA citation network. Path analysis has more recently been undertaken for documents in the area of hypertext research using author co-citations (Chen & Carr, 1999).

Taking citation links as the basis of a structural analysis of science, it is natural to suppose that it would be possible to travel from any topic or field to any other (Small, 1999) just as in the world of the Internet we might follow a series of hypertext links to reach any desired Web site. In the abstract, this is equivalent to traversing a network, but there is no guarantee the structure is in fact connected. In science, citations are very unevenly distributed, concentrating in narrowly defined pockets which correspond roughly to specialties or invisible colleges of researchers (Small & Griffith, 1974). The boundaries of these regions of high density are not well defined, however. Yet the most interesting links in the chain from one end of science to the other are those which cross disciplinary boundaries. Interdisciplinary links represent a kind of intellectual leap from one domain to another.

In the world of citation analysis, strong links can be established by frequent patterns of co-citation (Small, 1973) or bibliographic coupling (Kessler, 1963). Co-citation links are a second order form of citation linkage that depends on the joint citing of two earlier documents by later documents. Unlike direct citation links, co-citations are nondirectional and can be weighted by frequency of occurrence. By simple "thresholding," it is possible to identify regions of high co-citation density. Thresholding is in fact equivalent to the method of clustering called "single-linkage" (Hartigan, 1975).

In a map based on co-citation clusters, an interdisciplinary link can occur when an author co-cites across the boundary of two disciplinary clusters. If the author cites predominantly into one cluster, as is often the case, the interdisciplinary co-citation reaches out beyond the author's home cluster (see Figure 1). This reaching out or stretching can import or export methods, ideas, models, or empirical results from the author's field to the other field. This is an act requiring a broad awareness of literature plus the creative imagination to see how the outside information fits with the author's problem domain. The author of such a paper is going out on a limb to integrate ideas from another discipline.

The objective of the present study is to examine the nature of the connections that tie the scientific literature together, focusing particularly on links crossing disciplinary boundaries. The question is whether interdisciplinary transitions are gradual or abrupt or based on shared features, analogies, creative insights, or perhaps even questionable assumptions—in short, how far the author had to stretch to make the connection. In another sense it is an examination of the creative process of moving from one domain of knowledge to another. If citation relationships capture authors' decisions or selections on what documents are relevant to a problem, paths that follow citation links may in some sense capture steps in problem-solving behavior, logical thinking, or intuition.

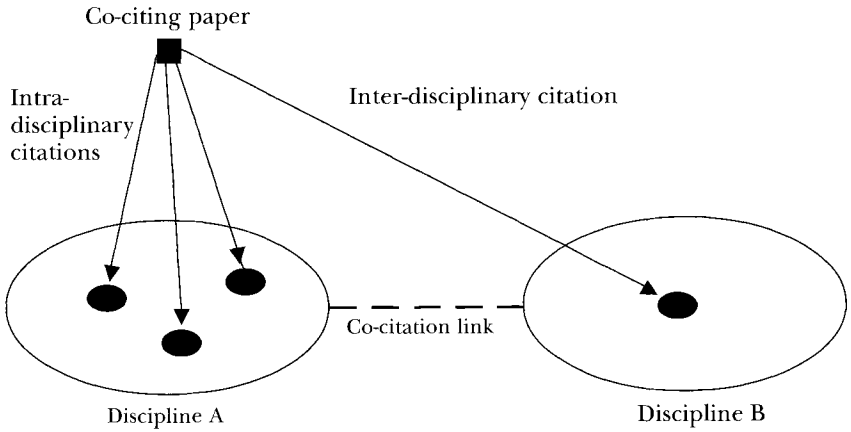


Figure 1. Cross-Disciplinary Citations. A schematic representation of citation patterns from a single citing paper (black square) to four prior cited documents (black ovals), three of which are situated in one research area and one in another. One of the citations crosses a disciplinary boundary thereby creating a link between the two disciplines.

Information Pathways

The basic requirement of a pathway through science is that the linked objects form a chain of significant connections. Ideally each connection represents a relationship whose logic can be determined by some form of content analysis. In an abstract sense, an information pathway could be defined as a sequence or succession of information objects or events (documents, descriptors, topics) such that each object along the path bears some kind of relationship to the objects that precede it. Of course, the zero order case is a random path or walk in which there is no relationship, or at least an arbitrary one, between successive objects. Order can be imposed on this succession by introducing various types of formal restrictions. For example, a path through scientific papers might be required to follow citation or co-citation links or other form of document association. Other types of restrictions are whether to allow the repetition of objects along the path, whether the path is exhaustive or complete—that is, all objects must be visited—or whether the path is to be as short as possible (Harary, 1972). Optimal paths of various kinds can be defined such as the shortest path that visits all nodes in the network, called the "traveling salesman" problem (Simon, 1969).

Due to the complexity of the citation graph, it is impossible to create a nonrepeating, linear path through all of science, touching all papers only once if the path is constrained to follow specific links. Only the

simplest of graph structures would allow this. However, nonrepeating sequences of objects can be formed by relaxing the requirement that each document must be linked to its predecessor. The depth- or breadth-first search techniques are effective when a complete tour of objects is desired that remains as coherent as possible (Sedgewick, 1983). The depth-first search was used to transform a co-citation graph for a cancer research area into a linear narrative (Small, 1986).

Linear nonrepetitive paths necessarily exist through a network connecting any two arbitrarily selected points, provided of course the graph is connected and contains no unreachable subcomponents (Hillier & Lieberman, 1967). This is the type of path illustrated below. The path should follow strong links but need not be the shortest path. Shortest paths are reminiscent of the small world experiments in sociology (Milgram, 1967; Garfield, 1981; Kochen, 1989) where the smallest number of intervening acquaintances between two arbitrarily selected individuals is sought. In the citation world, short paths might arise, for example, if a paper in astrophysics cites a paper in sociology. Such paths, though occasionally seen, are not the norm and usually are idiosyncratic and have low frequency. More interesting are paths that follow links established by multiple authors and hence represent a consensus or congruence of opinion. These might be called high frequency or well traveled paths.

Co-Citation Mapping Methods

The creation of pathways through science begins with a hierarchical clustering of highly cited papers in which co-citation serves as the measure of association between papers (Small, 1999). This is carried out in a series of iterations until as much as possible of the corpus of scientific literature can be amalgamated into a single hierarchical structure several levels deep. The number of levels required depends on the number of starting documents which in turn depends on the thresholds set for defining what is considered highly cited. A fractional citation counting method is used to ensure that papers are sampled across the various disciplines of science without biasing the selection to fields that inherently cite more than others (Small & Sweeney, 1985). In addition, an integer citation count threshold is used to avoid selection of infrequently cited papers.

In clustering an annual multidisciplinary database, several iterations of clustering are required to build up an overall structure. The output of each iteration becomes the input to the next, and residual co-citation links are recalculated to refer to the clustered objects at each step. At the end of the process, what is left are large-scale aggregates connected by rather weak links. These weak links make the macro-structure somewhat unstable over time, but they represent boundary-spanning events of considerable interest.

An integral part of the clustering process is creation of a spatial arrangement of the objects at each level or iteration. This is achieved by a geometric triangulation procedure that converts each link into a distance measure (Lee et al., 1977). The unit of distance is called the Garfield and is given by the formula:

Distance A-B = (1 – similarity)/(1 – similarity threshold)
Similarity = co-cites A-B/sqrt (cites A* cites B)

Note: When the similarity is equal to the similarity threshold, the distance is equal to one Garfield—the distance associated with the weakest link on the map.

The positioning of objects is accomplished by taking the two strongest links (shortest distances) for each object to be added to the map (Small, 1997). After each cluster is configured by triangulating on the strongest links, the structures are integrated hierarchically. This involves expanding the higher level objects and translating the coordinates of lower level objects so that they fit into them. In two dimensions, objects are represented as circles whether they are clusters or documents and, because the structure is hierarchical, the larger circles contain smaller circles, the smallest ones being the documents themselves. In three dimensions, circles become spheres containing smaller spheres.

The 1996 Map of Science

The procedure for generating the science map for 1996 was similar to that used for a 1995 map (Small, 1999). An integer citation threshold of six (6) and a fractional threshold of 1.0 were set to select papers cited in the 1996 *Science Citation Index®* (*SCI*) file. The cited references were restricted to publication dates in a fifteen-year period 1982 to 1996. Only the results for the main cluster hierarchy will be presented here. These are the papers included in the largest hierarchical grouping. Table 1 shows the number of clusters and documents for each of the five levels. Thus 39,964 highly cited documents are contained in 4,723 level 1 clusters containing two or more documents. The 4,723 clusters are in turn contained in 757 level 2 clusters, which aggregate to form 159 level 3 objects. These

Table 1.
Objects in the Main Cluster Hierarchy

Level	*Number of Clusters*	*Number of Objects*	*Mean Objects*	*Mean Documents*
1	4,723	39,964	8.5	8.5
2	757	4,723	6.2	52.8
3	159	757	4.8	251.3
4	43	159	3.7	929.4
5	1	43	43.0	39,964.0

form 43 level 4 clusters, which amalgamate to a single group of 43 level 4 clusters at level 5, including all lower level clusters.

The map of science for 1996 (see Figure 2) is a linked structure of disciplines and research areas similar to those obtained for earlier annual files of the *SCI*. The map is predominantly linear in its progression from social science, biomedicine, chemistry, to physics. The social science areas are situated at the lower right and physics areas are in the upper left, although there is no significance to this general orientation. Neuroscience is situated above psychology and economics, and above neuroscience is a large central biomedical region. To the left and closely allied with biomedicine is protein chemistry and above it general chemistry. Ecology is situated to the left of chemistry and geoscience is to its left. Geoscience is just below physics, and above physics are surface science, materials, and optics. Computer science is at the social science/medicine pole of the map, linked to imaging and neural networks.

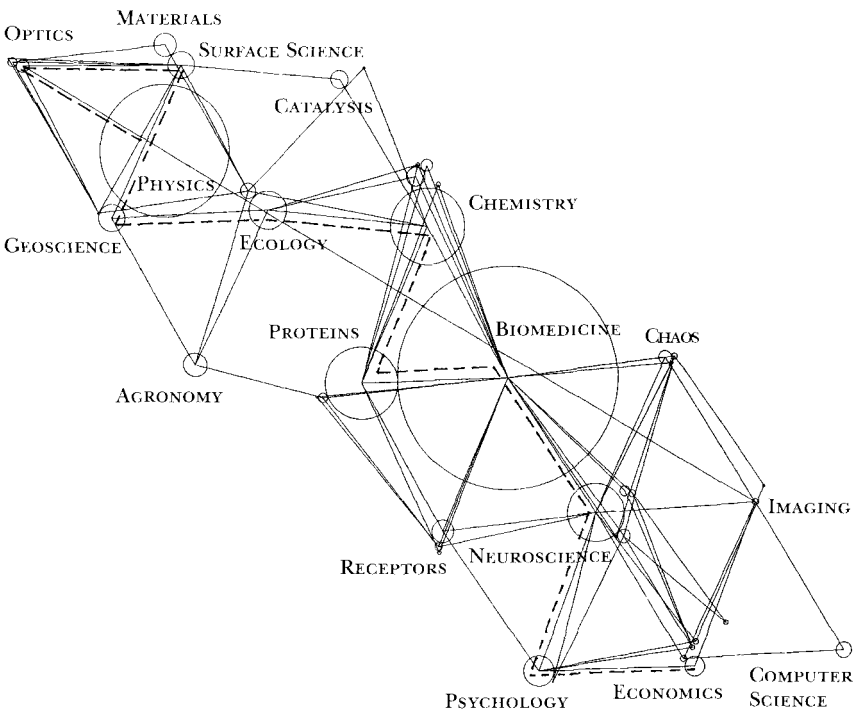


Figure 2. 1996 Map of Science. A network representation of the 43 fourth level clusters based on data from the 1996 *Science Citation Index* representing major scientific fields in social sciences, biomedicine, chemistry, and physics. Each high level cluster is represented by a circle and the links connecting them are aggregate document co-citations.

A Path from Economics to Physics

A path through science was generated by selecting starting and destination fields from among the forty-four high level clusters. To illustrate as wide a range of topics as possible, economics, shown at the lower right of Figure 2, was selected as the starting area and physics at the upper left was selected as the destination. Not only were these fields at opposite ends of the map, but they seemed at opposite intellectual poles—one in the worldly realm of human behavior the other in the extra-terrestrial. Specific papers within these regions were not specified, so the path algorithm was required only to find one beginning and one ending paper within each region.

The cluster hierarchy greatly simplifies finding strongly linked paths because the relatively few large-scale objects at the higher levels of aggregation can be traversed before descending to lower level objects, making the process one of gradually emerging detail and avoiding combinatorial complexity. The approach is to move down the hierarchy one level at a time. The path through the largest scale objects of course must begin and end with the starting and destination points. Among the objects at a given level, a minimal spanning tree is formed using the strongest co-citation links. A high level path is formed by navigating only those branches of the tree necessary to connect the starting and destination nodes. Then, moving down to the next level, for each successive pair of large-scale objects along this path, a pair of lower level objects is found that are most strongly linked by co-citation. This lower level pair thus links the larger objects. This defines starting and ending points within each large-scale object that can be navigated using the minimal spanning tree approach. Hence the process proceeds by alternating between finding paths through objects at some level and finding pairs of lower level objects spanning the higher-level

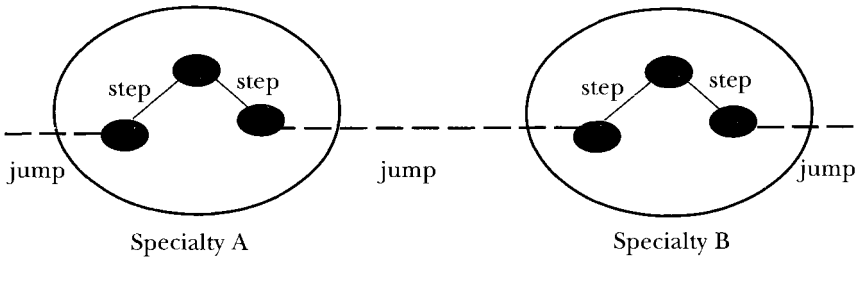


Figure 3. Path Formation: Steps and Jumps. A schematic representation of the two modes of traversal from node to node on the map. One mode, a step, follows a strong co-citation link between two objects that are contained in the same cluster. The other mode, a jump, follows a link between two objects not contained in the same cluster. Jumps tend to be weaker than steps.

object path. This continues until the document level is reached and a complete document pathway is formed.

Thus, the two kinds of path-forming processes are: (1) finding a sequence of lower level objects within a higher level object, and (2) finding the most strongly linked lower level objects within two adjacent higher-level objects. The two modes of traversal might be termed stepping and jumping, the one akin to stepping from stone to stone along a garden path, the other like jumping from one path to another. These are illustrated in Figure 3. The "jumps," of course, can involve traversing more weakly connected or distant objects and may entail larger shifts in subject matter while the "steps" are more strongly linked and closer in topic.

Table 2 gives the number of objects from each level touched by the completed path from economics to physics. Clearly the number of objects in the path decreases as the level increases, while the percentage of total objects increases. A total of 331 documents make up the final path corresponding to 330 transitions (either steps or jumps) from beginning to end. Table 3 shows how the transitions from object to object are distributed for each of the levels. For the document and first levels, there are about two steps for every jump. For higher levels (except the highest) there are about equal numbers of jumps and steps. Note that the sum of jumps and steps for a given level must equal the number of jumps at the next lower level.

Table 2.
OBJECTS IN THE PATH

Level	*Number of Objects*	*Percentage of Total*
Docs	331	0.8
1	121	2.6
2	42	5.5
3	23	14.5
4	11	25.6

It is also possible to compute the mean document co-citation strengths for transitions from object to object. The last column of Table 3 shows these mean values for steps (excluding jumps) between objects of each level. Mean co-citation strength diminishes with increasing level (except for the third and fourth levels), indicating that the larger aggregates are bound by weaker document co-citation links than the smaller aggregates. This is to be expected since ties are stronger at the local level.

DESCRIPTION OF THE PATH FROM ECONOMICS TO ASTROPHYSICS

The dotted line on Figure 2 shows how the path connects the highest level objects on the map of science. The path traverses the broad areas of

Table 3.
PATH STATISTICS

Level *Document*	*Jumps*	*Steps*	*Transitions*	*Objects*	*Mean Co-citations*
	120	210	330	331	14.7
1	41	79	120	121	11.5
2	22	19	41	42	10.8
3	10	12	22	23	6.0
4	0	10	10	11	7.9

economics, psychology, neuroscience, biomedicine, proteins, chemistry, ecology, geoscience, surface science, optics, and physics. Within economics, the starting point is a paper entitled "Making Fast Strategic Decisions in High-Velocity Environments." From here the path makes a transition from economics to psychology and moves into the psychology of work teams. The last paper in the path is a physics paper entitled "Wave Function of the Universe." Just prior to reaching this physics destination, the path traverses the topic of quantum field theory. The 331 documents comprising the full path are given in the Appendix. Headings interspersed in this list show the major subdivisions by cluster and subcluster, and the indentation of the heading indicates the hierarchical level of the subdivision.

Another way to view the progression from economics to physics is to plot co-citation frequency as a function of the position on the path (Figure 4). The number of co-citations is counted for each successive pair of documents. The figure is labeled with the subject matter of the highest level objects, indicating the points of transition between each by arrows along the vertical axis. The highest co-citation rates are concentrated in biomedicine, neuroscience, and proteins. High co-citation rates are sometimes correlated with subject matter. For example, within neuroscience, high rates are observed in the section dealing with nitric oxide as a neuronal messenger, and in surface science, for the topic of quantum dots. The steep spikes occasionally emerging above the terrain indicate the linking of highly cited technique papers.

A detailed interpretation of the entire path will not be attempted here. However, discussion of fifty or so documents comprising the psychology section should suffice to give a general flavor. The starting paper on decision making leads to the psychology of work teams and team leadership. Work teams then move to the more general concept of groups and differences in male and female participation. This leads to gender stereotypes and a focus on category-based versus individual impression formation. The role of memory in judgment emerges from this and how expectations interact with memory. Memory of persons gives way to attribution of cause and effect and causal thinking. The role of positive and

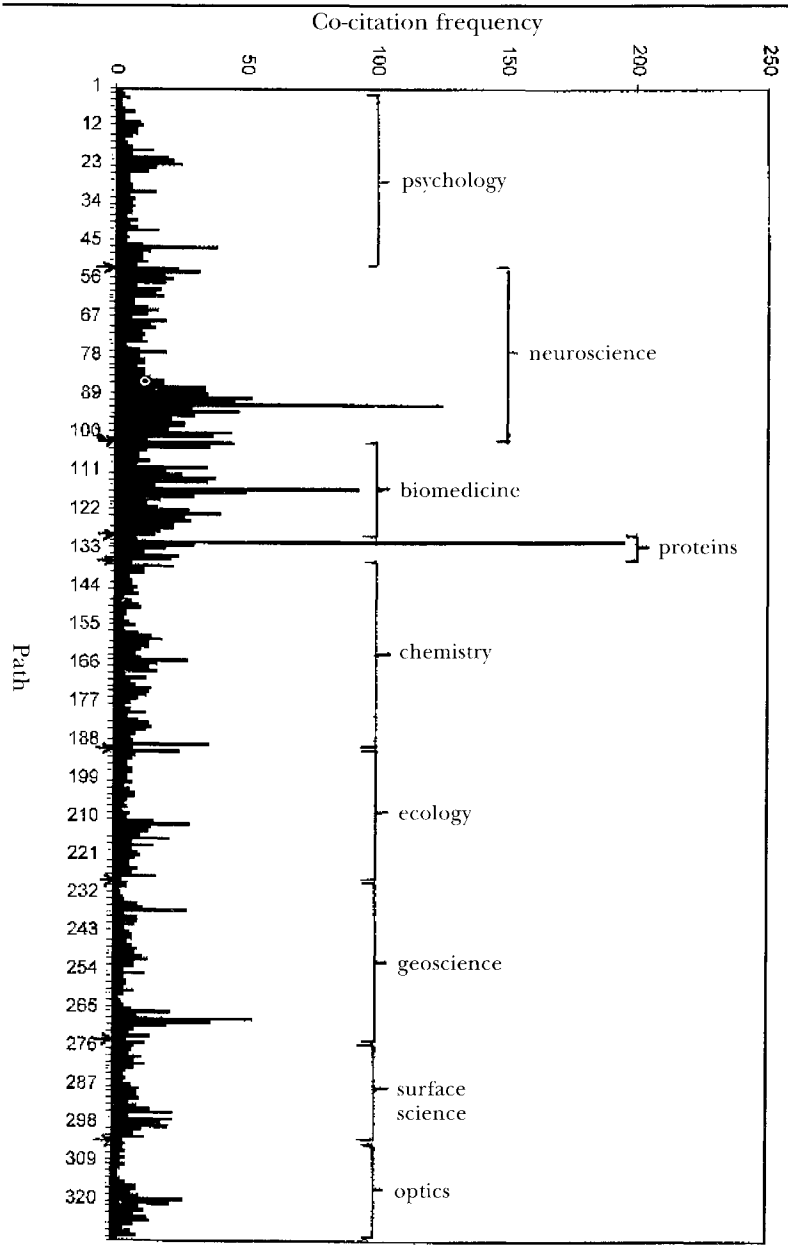


Figure 4. Co-citation Density along Path. The graph plots the co-citation frequencies of adjacent papers along the path from economics to physics. There are 331 papers and thus 330 links connecting them. The ranges of papers in major disciplines are indicated by brackets, and arrows on the vertical axis indicate the points of transition between disciplines.

negative events on thinking leads to perceptions of one's own health. Negative feelings on health then progress to measuring affect, and in turn to scales for assessing depression. Measuring depression links to treatments for depression and recurrent depression. Rapid cycling emerges from recurrent depression, then progressing to bipolar disorder. Expressed emotion is a common diagnostic method for both bipolar disorder and schizophrenia, which is the next topic. Drug treatment of schizophrenia moves to studies of cerebral blood flow, eventually ending up in the field of neuroscience with the study of the normal brain.

The above description exemplifies many of the types of shifts seen in the path as a whole: movement from the specific to the general, from the normal to the abnormal, from the aggregate to the disaggregate, the macro to the micro, and so on. But these shifts do not tell the whole story. For example, the gradual transition from studies of thinking to studies of depression is particularly striking. It appears to be mediated by the perception of well-being and specifically negative thinking and emotion regarding one's health. Health complaints are, in effect, generalized to adverse emotional states that then progress to depression or anxiety.

Table 4 summarizes the path in terms of the main subtopics visited and the corresponding range of documents traversed. The subtopic list corresponds to divisions at clustering levels below the fourth level, and the intent was to break the path into segments of approximately equal length. The document ranges correspond to the numbering scheme in the Appendix. The approximate nature of the boundaries established by clustering are evident in the case of geoscience and its transition to surface science. Some solid-state topics, such as the equation of state for solids and ab-initio molecular dynamics, have been incorporated into geoscience that might more logically have been assigned to surface science. Despite this boundary question, the topics progress from one to another in a logical and regular manner.

The sequence of research areas reflects a plausible scenario for interweaving the scientific fabric. From the broadest perspective there is the progression from human to biological to physical sciences. Moving down a level, there is a progression of general themes. Starting with social groups, the progression is to individual behavior. Normal behavior leads to considerations of abnormal behavior. In trying to understand the basic biology of abnormal behavior, the path leads back to study of the normal brain in neuroscience. Within the nervous system, the level of the neuron is reached, at which point chemical neurotransmitters enter the picture. This biochemical level moves from the nervous system to the immune system and immune disease. The attempt to understand the biochemistry of AIDS leads to structural analysis of biochemical molecules and to chemistry in general. The traversal of chemistry follows catalytic processes and ends up with photosynthesis, which leads in turn to

Table 4.
TOPICS ALONG THE PATH

Level 4 Cluster	*Subtopic (Levels 3 – 1)*	*Document Range*
Economics	Decision making	1
Psychology	Social groups	2 – 17
	Well-being	18 – 25
	Depression	26 – 37
	Expressed emotion	38 – 42
	Schizophrenia	43 – 51
Neuroscience	Visual cortex	52 – 65
	Thalamic neurons	67 – 72
	Synaptic transmission	78 – 86
	Nitric-oxide messenger	88 – 101
Biomedicine	Nuclear factor kappa-B	102 – 104
	Tumor necrosis factor	105 – 108
	Interleukin	109 – 114
	Drug resistant HIV	115 – 128
Proteins	Protein structure	130 – 135
Chemistry	Metallocene catalysis	137 – 140
	Khand reaction	141 – 146
	Palladium catalysis	147 – 154
	Asymmetric catalysis	156 – 168
	Porphyrins	169 – 189
Ecology	Atmospheric CO_2	193 – 216
	Models of biosphere	217 – 227
Geoscience	Climate cycles	228 – 237
	Earth's geoid	238 – 248
	Earth's mantle	250 – 253
	Seismic velocity	256 – 258
	Equation of state for solids	263 – 266
	Ab-initio molecular dynamics	269 – 273
Surface science	Diamond surface	274 – 281
	Epitaxial surface growth	282 – 295
	Quantum dots	296 – 302
Optics	Quantum wells	303 – 322
	Quantum field theory	323 – 330
Physics	Astrophysics	331

considerations of the earth's atmosphere. Climate studies follow, and we land back on the earth's crust and the physical processes that govern the earth's mantle. Moving down to the atomic level for understanding solids, the focus turns to the surfaces of solids. Descending to yet finer scales, quantum phenomena are encountered and their abstract mathematical treatment, leading finally to theories of the universe as a whole within the field of astrophysics.

The most striking aspect of this broad brush discussion of main themes is how the focus alternates from large to small scale, from the group to the individual, from diseases to molecules, from molecules to the atmosphere, from the earth back down to the atomic and quantum levels, and finally

up again to the universe. But, to understand the logic behind these shifts in topic, it is necessary to consider transitions at the document to document level.

Understanding Transitions between Disciplines

A proper analysis of the nature of the topic transitions would require a content analysis of the co-citation passages for pairs of documents along the path. However, a first order approximation to understanding the nature of the transitions can be based on a close examination of the titles of the linked documents. Many of the transitions resemble what might be called substitution around a stable point of reference in which one aspect or theme changes while another remains constant. The following cases illustrate this principle using some of the main disciplinary transitions along the path.

The transition from psychology to neuroscience starts with the study of patterns of cerebral blood flow in schizophrenia using positron emission tomography (PET). PET can also be used to study willed action and word usage in the normal prefrontal cortex. Schizophrenia then leads to the study of the normal human cortex via the reference point provided by the PET technique, and a normal activity is substituted for the abnormal condition.

The transition from neuroscience to immunology involves going from the neuronal messenger nitric oxide to HIV. The common link is a substance that appears to play a role in controlling their biochemistry. Nitric oxide in the nervous system is synthesized by the nitric oxide synthase gene, and the encoding of the gene is induced by a substance called nuclear factor kappa-B. This substance also plays a role in the immune system where it can activate HIV. Thus the topic transition occurs around a point of reference provided by the substance nuclear factor Kappa-B with HIV activation being substituted for the stimulation of the nitric-oxide gene.

In the transition from biomedicine to biochemistry, the topic moves from the study of biologically important proteins, in particular DNA-polymerase, to the use of computer plotting methods, such as molscript or electron density mapping, to study protein structure. Here the reference point is the protein molecule, while biological function of the molecule is exchanged for the physical depiction of its structure.

The transition from protein structure to chemistry proper is more subtle but seems to hinge on the structural specificity of catalytic reactions. The study of molecular structure by x-ray diffraction thus leads to the mechanism of catalytic reactions. The point of reference is molecular structure itself, and the study of catalytic reactions replaces the methodology of x-ray diffraction.

The transition from chemistry to ecology is simpler and involves going from the study of artificial photosynthesis to natural photosynthesis in

bacteria, photosynthesis being the common thread. From ecology to geoscience involves going from the development of meteorological models of general atmospheric circulation to the study of climate changes over long periods of time. Climate is the common thread, and historical cycles substitute for model building. Hence, each of the interdisciplinary transitions seems to involve a common thread and a substitution.

Conclusion

Examination of these major topic shifts suggests that some general principles may be at work. Since the transitions are brought about by the behavior of the co-citing authors, understanding these patterns will ultimately involve an examination of creative information seeking by authors and, in particular, how authors in one field reach out for information in another field. Based on this preliminary content analysis, it is possible to identify some possible mechanisms and strategies authors use to bridge information gaps.

First, there is the mechanism of extending a topic with a gradual shift in its scope. For example, in psychology, negative affect was extended to depression and hence to rapid mood cycling and bipolar disorder. Extending can also be literal, importing a concept or method into another domain without modification, as in the case of PET in the transition from psychology to neuroscience. Extension provides the common point of reference in many of the interdisciplinary transitions. The second mechanism is the substitution of one entity for another. This was seen in many of the cross-disciplinary transitions—e.g., the HIV substitution for nitric oxide in going from neuroscience to biomedicine.

The joint operation of extension and substitution might be seen as a simple kind of progression by analogy in which "*A* is to *B* as *B* is to *C*" where *B* is the common thread or point of reference. True analogies of the form "*A* is to *B* as *C* is to *D*" are also possible and could emerge, for example, if an extension of topic *B* transformed it into a distinct entity *D*. True analogies are also more tenuous than extension with substitution. It is tempting to postulate that the author first sees the possibility of a link up with another domain as a true analogy but, as his or her thinking firms up, there is a realization that *B* and *D* can be made equivalent in some sense, perhaps by transforming one into the other by logical extension. Thus the concrete transition emerges from the initial glimmer of an analogy.

It seems plausible that the creative use of information involves some form of thinking by analogy and the recognition of similar structures in disparate domains. Returning to the retrieval example at the beginning of this article, it would be as if we were to take two apparently unrelated items from a search output and ask subjects to think of ways the two items might be related or brought into a common framework. If a path-finding

algorithm were in place within the retrieval system, then a document path between them could be generated. This might reveal new ways that seemingly unrelated pieces of information can be related and thereby provide creative insights, new hypotheses, or perhaps even aid in discovery. It also suggests that the ability of users to see the relevance of apparently nonrelevant pieces of information is a key step in the discovery process. Path-finding techniques might offer a new approach to revealing such hidden or potential relevance.

Epistemology and the Unity of Science

In addition to information discovery, path creation might have application to the evaluation of information and to an epistemological warrant for science. There appears to be a movement in philosophy toward making scientific belief more a function of group than individual cognition (Schmitt, 1994). It seems reasonable to postulate that the unity and coherence of science, and therefore the existence of pathways, is related to the solidity of the scientific findings (Small, 1998). This is based on the notion that of all the documents vying for attention, the most promising ones are those most closely tied to the existing body of strongly verified knowledge (Stent, 1972). Assuming that it were possible to identify a core of strongly verified documents in science, an attempt could be made to connect any new document to the verified core by a pathfinding algorithm.

A text with connections to strongly confirmed or verified knowledge would deserve more serious attention than one without such connections. Texts lacking connections would be treated with greater caution and skepticism. Topic connections may offer a way of evaluating uncontrolled information sources on the Web (Kleinberg, 1998). Grafton (1997) points out, however, the mere existence of linkages or the quoting of sources does not guarantee truth or objectivity. Thus the nature as well as number of links is critical, and the existence of a path could only be considered an indicator, not an infallible guide.

While this article has focused on linear paths that connect arbitrarily selected beginning and ending documents, other kinds of paths may also be of interest, particularly those that provide a complete tour of the network. Complete paths, if properly constructed, could provide a comprehensive review of a subject area and ultimately an excursion through the entire fabric of science. A promising approach for achieving this is the identification of the longest linear routes through the minimal spanning tree representation for each cluster, to preserve, as far as possible, a coherent sequential flow of ideas. This could be coupled with a breadth-first search on the tree to explore side branches, much as one would digress from a main topic. Since the structure is hierarchical, this process could progress down the hierarchy until the document level is reached,

piecing together the document sequences like strands of DNA. The final result would be a linear ordering of all the documents in the structure, a kind of complete genome sequence of science.

REFERENCES

Bush, V. (1945). As we may think. *Atlantic Monthly, 176*(1), 101-108.

Chen, C., & Carr, L. (1999). Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998). In *Proceedings of the 10th ACM Conference on Hypertext* (Hypertext '99, February 21-25, 1999. Darmstadt, Germany) (pp. 51-60). New York: ACM Press.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities.* Chicago, IL: University of Chicago Press.

Garfield, E. (1981). It's a small world after all. In E. Garfield (Ed.), *Essays of an information scientist* (vol. 4, pp. 299-304). Philadelphia, PA: ISI Press.

Grafton, A. (1997). *The footnote: A curious history* (rev. ed.). Cambridge, MA: Harvard University Press.

Harary, F. (1972). *Graph theory.* Reading, MA: Addison-Wesley.

Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

Hillier, F. S., & Lieberman, G. J. (1967). *Introduction to operations research.* San Francisco, CA: Holden-Day.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11,* 39-63.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10-25.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In H. Karloff (Ed.), *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.* New York: ACM Press.

Kochen, M. (Ed). (1989). *The small world.* Norwood, NJ: Ablex.

Kuhlthau, C. C. (1999). Accommodating the user's information search process: Challenges for information retrieval system designers. *Bulletin of the American Society for Information Science, 25*(3), 12-16.

Lee, R. C. T.; Slagle, J. R.; & Blum, H. (1977). A triangulation method for the sequential mapping of points from *N*-space to two-space. *IEEE Transactions on Computers, 26*(1), 288-292.

Lemaine, G.; MaCleod, R.; Mulkay, M..; & Weingart, P. (Eds.). (1976). *Perspectives on the emergence of scientific disciplines.* The Hague, Netherlands: Mouton.

Milgram, S. (1967). The small-world problem. *Psychology Today, 1*(1), 61-67.

Neurath, O. (1938). Unified science as encyclopedic integration. In O. Neurath, R. Carnap, & C. Morris (Eds.), *Foundations of the unity of science: Toward an international encyclopedia of unified science* (vol. 1, pp. 1-27). Chicago, IL: University of Chicago Press.

Schmitt, F. F. (Ed.). (1994). *Socializing epistemology: The social dimensions of knowledge.* Lanham, MD: Rowman & Littlefield.

Sedgewick, R. (1983). *Algorithms.* Reading, MA: Addison-Wesley.

Simon, H. A. (1969). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies, 4*(1), 17-40.

Small, H., & Sweeney, E. (1985). Clustering the *Science Citation Index*® using co-citations, I: A comparision of methods. *Scientometrics, 7*(3-6), 391-409.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science, 37*(3), 97-110.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics, 38*(2), 275-293.

Small, H. (1998). Citations and consilience in science. *Scientometrics, 43*(1), 143-148.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Stent, G. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American, 227*(6), 84-93.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*(2), 183-203.

White, H., & McCain, K. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32,* 3-72.

Wilson, E. O. (1998). *Consilience: The unity of knowledge.* New York: Knopf.

APPENDIX

DOCUMENT PATH FROM ECONOMICS TO PHYSICS

ECONOMICS

1 EISENHARDT KM,ACAD MGMT J,vol 0032,page 0543,1989,cites= 33,MAKING FAST STRATEGIC DECISIONS INHIGH-VELOCITY ENVIRONMENTS

PSYCHOLOGY

SOCIAL GROUPS

2 GERSICK CJG,ACAD MGMT J,vol 0031,page 0009,1988,cites= 12,TIME AND TRANSITION IN WORK TEAMS - TOWARD A NEW MODEL OF GROUP DEVELOPMENT
3 SUNDSTROM E,AM PSYCHOL,vol 0045,page 0120,1990,cites= 17,WORK TEAMS - APPLICATIONS AND EFFECTIVENESS
4 MANZ CC,ADM SCI QUA,vol 0032,page 0106,1987,cites= 11,LEADING WORKERS TO LEAD THEMSELVES - THE EXTERNAL LEADERSHIP OF SELF-MANAGING WORK TEAMS
5 WHEELAN SA,SEX ROLES,vol 0027,page 0001,1992,cites= 6,DIFFERENCES IN MALE AND FEMALE PATTERNS OF COMMUNICATION IN GROUPS - A METHODOLOGICAL ARTIFACT
6 WOOD W,PSYCHOL B,vol 0102,page 0053,1987,cites= 8,META-ANALYTIC REVIEW OF SEX-DIFFERENCES IN GROUP-PERFORMANCE
7 EAGLY AH,J PERS SOC,vol 0060,page 0685,1991,cites= 15,GENDER AND THE EMERGENCE OF LEADERS - A METAANALYSIS
8 EAGLY AH,PSYCHOL B,vol 0108,page 0233,1990,cites= 23,GENDER AND LEADERSHIP-STYLE - A METAANALYSIS
9 EAGLY AH,J PERS SOC,vol 0046,page 0735,1984,cites= 17,GENDER STEREOTYPES STEM FROM THE DISTRIBUTION OF WOMEN AND MEN INTO SOCIAL ROLES
10 FISKE ST,ADV EXP SOC,vol 0023,page 0001,1990,cites= 53,A CONTINUUM OF IMPRESSION-FORMATION, FROM CATEGORY-BASED TO INDIVIDUATING PROCESSES - INFLUENCES OF INFORMATION AND MOTIVATION ON ATTENTION AND INTERPRETATION
11 SRULL TK,PSYCHOL REV,vol 0096,page 0058,1989,cites= 30,PERSON MEMORY AND JUDGMENT
12 STANGOR C,PSYCHOL B,vol 0111,page 0042,1992,cites= 20,MEMORY FOR EXPECTANCY-CONGRUENT AND EXPECTANCY-INCONGRUENT INFORMATION - A REVIEW OF THE SOCIAL AND SOCIAL DEVELOPMENTAL LITERATURES
13 SRULL TK,J EXP PSY L,vol 0011,page 0316,1985,cites= 14,ASSOCIATIVE STORAGE AND RETRIEVAL-PROCESSES IN PERSON MEMORY
14 HASTIE R,J PERS SOC,vol 0046,page 0044,1984,cites= 19,CAUSES AND EFFECTS OF CAUSAL ATTRIBUTION
15 WEINER B,PSYCHOL B,vol 0097,page 0074,1985,cites= 15,SPONTANEOUS CAUSAL THINKING
16 ROESE NJ,J PERS SOC,vol 0066,page 0805,1994,cites= 7,THE FUNCTIONAL BASIS OF COUNTERFACTUALTHINKING
17 TAYLOR SE,PSYCHOL B,vol 0110,page 0067,1991,cites= 29,ASYMMETRICAL EFFECTS OF POSITIVE AND NEGATIVE EVENTS - THE MOBILIZATION MINIMIZATION HYPOTHESIS

WELL-BEING

18 TAYLOR SE,PSYCHOL B,vol 0103,page 0193,1988,cites= 114,ILLUSION AND WELL-BEING - A SOCIAL PSYCHOLOGICAL PERSPECTIVE ON MENTAL-HEALTH
19 SCHEIER MF,HEALTH PSYC,vol 0004,page 0219,1985,cites= 78,OPTIMISM, COPING, AND HEALTH - ASSESSMENT AND IMPLICATIONS OF GENERALIZED OUTCOME EXPECTANCIES

20 WATSON D,PSYCHOL REV,vol 0096,page 0234,1989,cites= 77,HEALTH COMPLAINTS, STRESS, AND DISTRESS - EXPLORING THE CENTRAL ROLE OF NEGATIVE AFFECTIVITY
21 WATSON D,PSYCHOL B,vol 0096,page 0465,1984,cites= 97,NEGATIVE AFFECTIVITY - THE DISPOSITION TO EXPERIENCE AVERSIVE EMOTIONAL STATES
22 WATSON D,PSYCHOL B,vol 0098,page 0219,1985,cites= 80,TOWARD A CONSENSUAL STRUCTURE OF MOOD
23 WATSON D,J PERS SOC,vol 0054,page 1063,1988,cites= 116,DEVELOPMENT AND VALIDATION OF BRIEF MEASURES OF POSITIVE AND NEGATIVE AFFECT - THE PANAS SCALES
24 CLARK LA,J ABN PSYCH,vol 0100,page 0316,1991,cites= 41,TRIPARTITE MODEL OF ANXIETY AND DEPRESSION - PSYCHOMETRIC EVIDENCE AND TAXONOMIC IMPLICATIONS
25 BECK AT,CLIN PSYCH,vol 0008,page 0077,1988,cites= 134,PSYCHOMETRIC PROPERTIES OF THE BECK DEPRESSION INVENTORY - 25 YEARS OF EVALUATION

DEPRESSION

26 ELKIN I,ARCH G PSYC,vol 0046,page 0971,1989,cites= 75,NATIONAL-INSTITUTE-OF-MENTAL-HEALTH TREATMENT OF DEPRESSION COLLABORATIVE RESEARCH-PROGRAM - GENERAL EFFECTIVENESS OF TREATMENTS
27 BELSHER G,PSYCHOL B,vol 0104,page 0084,1988,cites= 13,RELAPSE AFTER RECOVERY FROM UNIPOLAR DEPRESSION - A CRITICAL-REVIEW
28 KELLER MB,J AM MED A,vol 0250,page 3299,1983,cites= 15,PREDICTORS OF RELAPSE IN MAJOR DEPRESSIVE DISORDER
29 KELLER MB,ARCH G PSYC,vol 0049,page 0809,1992,cites= 31,TIME TO RECOVERY, CHRONICITY, AND LEVELS OF PSYCHOPATHOLOGY IN MAJOR DEPRESSION - A 5-YEAR PROSPECTIVE FOLLOW-UP OF 431 SUBJECTS
30 FRANK E,ARCH G PSYC,vol 0047,page 1093,1990,cites= 46,3-YEAR OUTCOMES FOR MAINTENANCE THERAPIES IN RECURRENT DEPRESSION
31 PRIEN RF,ARCH G PSYC,vol 0041,page 1096,1984,cites= 33,DRUG-THERAPY IN THE PREVENTION OF RECURRENCES IN UNIPOLAR AND BIPOLAR AFFECTIVE-DISORDERS - REPORT OF THE NIMH COLLABORATIVE STUDY-GROUP COMPARING LITHIUM-CARBONATE, IMIPRAMINE, AND A LITHIUM-CARBONATE IMIPRAMINE COMBINATION
32 CORYELL W,ARCH G PSYC,vol 0049,page 0126,1992,cites= 13,RAPIDLY CYCLING AFFECTIVE-DISORDER - DEMOGRAPHICS, DIAGNOSIS, FAMILY HISTORY, AND COURSE
33 BAUER MS,AM J PSYCHI,vol 0151,page 0506,1994,cites= 13,MULTISITE DATA REANALYSIS OF THE VALIDITY OF RAPID-CYCLING AS A COURSE MODIFIER FOR BIPOLAR DISORDER IN DSM-IV
34 WEHR TA,AM J PSYCHI,vol 0144,page 1403,1987,cites= 19,CAN ANTIDEPRESSANTS CAUSE MANIA AND WORSEN THE COURSE OF AFFECTIVE-ILLNESS
35 PEET M,BR J PSYCHI,vol 0164,page 0549,1994,cites= 14,INDUCTION OF MANIA WITH SELECTIVE SEROTONIN REUPTAKE INHIBITORS AND TRICYCLIC ANTIDEPRESSANTS
36 SACHS GS,J CLIN PSY,vol 0055,page 0391,1994,cites= 15,A DOUBLE-BLIND TRIAL OF BUPROPION VERSUS DESIPRAMINE FOR BIPOLAR DEPRESSION
37 GELENBERG AJ,N ENG J MED,vol 0321,page 1489,1989,cites= 22,COMPARISON OF STANDARD AND LOW SERUM LEVELS OF LITHIUM FOR MAINTENANCE TREATMENT OF BIPOLAR DISORDER

EXPRESSED EMOTION

38 MIKLOWITZ DJ,ARCH G PSYC,vol 0045,page 0225,1988,cites= 20,FAMILY FACTORS AND THE COURSE OF BIPOLAR AFFECTIVE-DISORDER

39 HOOLEY JM,BR J PSYCHI,vol 0148,page 0642,1986,cites= 18,LEVELS OF EXPRESSED EMOTION AND RELAPSE IN DEPRESSED-PATIENTS
40 KAVANAGH DJ,BR J PSYCHI,vol 0160,page 0601,1992,cites= 32,RECENT DEVELOPMENTS IN EXPRESSED EMOTION AND SCHIZOPHRENIA
41 TARRIER N,BR J PSYCHI,vol 0153,page 0532,1988,cites= 22,THE COMMUNITY MANAGEMENT OF SCHIZOPHRENIA - A CONTROLLED TRIAL OF A BEHAVIORAL INTERVENTION WITH FAMILIES TO REDUCE RELAPSE
42 HOGARTY GE,ARCH G PSYC,vol 0043,page 0633,1986,cites= 36,FAMILY PSYCHOEDUCATION, SOCIAL SKILLS TRAINING, AND MAINTENANCE CHEMOTHERAPY IN THE AFTERCARE TREATMENT OF SCHIZOPHRENIA .1. ONE-YEAR EFFECTS OF A CONTROLLED-STUDY ON RELAPSE AND EXPRESSED EMOTION

SCHIZOPHRENIA

43 HOGARTY GE,ARCH G PSYC,vol 0045,page 0797,1988,cites= 18,DOSE OF FLUPHENAZINE, FAMILIALEXPRESSED EMOTION, AND OUTCOME IN SCHIZOPHRENIA - RESULTS OF A 2-YEAR CONTROLLED-STUDY
44 MARDER SR,ARCH G PSYC,vol 0044,page 0518,1987,cites= 14,LOW-DOSE AND CONVENTIONAL-DOSE MAINTENANCE THERAPY WITH FLUPHENAZINE DECANOATE - 2-YEAR OUTCOME
45 BALDESSARINI RJ,ARCH G PSYC,vol 0045,page 0079,1988,cites= 34,SIGNIFICANCE OF NEUROLEPTIC DOSE AND PLASMA-LEVEL IN THE PHARMACOLOGICAL TREATMENT OF PSYCHOSES
46 MARDER SR,AM J PSYCHI,vol 0151,page 0825,1994,cites= 92,RISPERIDONE IN THE TREATMENT OF SCHIZOPHRENIA
47 KANE J,ARCH G PSYC,vol 0045,page 0789,1988,cites= 184,CLOZAPINE FOR THE TREATMENT-RESISTANT SCHIZOPHRENIC - A DOUBLE-BLIND COMPARISON WITH CHLORPROMAZINE
48 CARPENTER WT,AM J PSYCHI,vol 0145,page 0578,1988,cites= 52,DEFICIT AND NONDEFICIT FORMS OF SCHIZOPHRENIA - THE CONCEPT
49 MCGLASHAN TH,ARCH G PSYC,vol 0049,page 0063,1992,cites= 20,THE POSITIVE-NEGATIVE DISTINCTION IN SCHIZOPHRENIA - REVIEW OF NATURAL-HISTORY VALIDATORS
50 LIDDLE PF,BR J PSYCHI,vol 0151,page 0145,1987,cites= 47,THE SYMPTOMS OF CHRONIC-SCHIZOPHRENIA - A RE-EXAMINATION OF THE POSITIVE-NEGATIVE DICHOTOMY
51 LIDDLE PF,BR J PSYCHI,vol 0160,page 0179,1992,cites= 67,PATTERNS OF CEREBRAL BLOOD-FLOW IN SCHIZOPHRENIA NEUROSCIENCE

VISUAL CORTEX

52 FRITH CD,P ROY SOC B,vol 0244,page 0241,1991,cites= 58,WILLED ACTION AND THE PREFRONTAL CORTEX IN MAN - A STUDY WITH PET
53 PETERSEN SE,NATURE,vol 0331,page 0585,1988,cites= 108,POSITRON EMISSION TOMOGRAPHIC STUDIES OF THE CORTICAL ANATOMY OF SINGLE-WORD PROCESSING
54 FRISTON KJ,J CEREBR B,vol 0011,page 0690,1991,cites= 123,COMPARING FUNCTIONAL %PET< IMAGES - THE ASSESSMENT OF SIGNIFICANT CHANGE
55 WATSON JDG,CEREB CORT,vol 0003,page 0079,1993,cites= 53,AREA-V5 OF THE HUMAN BRAIN - EVIDENCE FROM A COMBINED STUDY USING POSITRON EMISSION TOMOGRAPHY AND MAGNETIC-RESONANCE-
IMAGING
56 ZEKI S,J NEUROSC,vol 0011,page 0641,1991,cites= 55,A DIRECT DEMONSTRATION OF FUNCTIONAL SPECIALIZATION IN HUMAN VISUAL-CORTEX
57 CORBETTA M,J NEUROSC,vol 0011,page 2383,1991,cites= 57,SELECTIVE AND DIVIDED ATTENTION DURING VISUAL DISCRIMINATIONS OF SHAPE, COLOR,

AND SPEED - FUNCTIONAL-ANATOMY BY POSITRON EMISSION TOMOGRAPHY
58 ZEKI S,NATURE,vol 0335,page 0311,1988,cites= 44,THE FUNCTIONAL LOGIC OF CORTICAL CONNECTIONS
59 LIVINGSTONE M,SCIENCE,vol 0240,page 0740,1988,cites= 84,SEGREGATION OF FORM, COLOR, MOVEMENT, AND DEPTH - ANATOMY, PHYSIOLOGY, AND PERCEPTION
60 LIVINGSTONE MS,J NEUROSC,vol 0004,page 0309,1984,cites= 51,ANATOMY AND PHYSIOLOGY OF A COLOR SYSTEM IN THE PRIMATE VISUAL-CORTEX
61 TSO DY,J NEUROSC,vol 0008,page 1712,1988,cites= 29,THE ORGANIZATION OF CHROMATIC AND SPATIAL INTERACTIONS IN THE PRIMATE STRIATE CORTEX
62 LIVINGSTONE MS,J NEUROSC,vol 0004,page 2830,1984,cites= 15,SPECIFICITY OF INTRINSIC CONNECTIONS IN PRIMATE PRIMARY VISUAL-CORTEX
63 MCGUIRE BA,J COMP NEUR,vol 0305,page 0370,1991,cites= 26,TARGETS OF HORIZONTAL CONNECTIONS IN MACAQUE PRIMARY VISUAL-CORTEX
64 GILBERT CD,J NEUROSC,vol 0009,page 2432,1989,cites= 39,COLUMNAR SPECIFICITY OF INTRINSIC HORIZONTAL AND CORTICOCORTICAL CONNECTIONS IN CAT VISUAL-CORTEX
65 TSO DY,J NEUROSC,vol 0006,page 1160,1986,cites= 39,RELATIONSHIPS BETWEEN HORIZONTAL INTERACTIONS AND FUNCTIONAL ARCHITECTURE IN CAT STRIATE CORTEX AS REVEALED BY CROSS-CORRELATION ANALYSIS
66 GRAY CM,NATURE,vol 0338,page 0334,1989,cites= 90,OSCILLATORY RESPONSES IN CAT VISUAL-CORTEX EXHIBIT INTER-COLUMNAR SYNCHRONIZATION WHICH REFLECTS GLOBAL STIMULUS PROPERTIES

THALAMIC NEURONS

67 STERIADE M,SCIENCE,vol 0262,page 0679,1993,cites= 86,THALAMOCORTICAL OSCILLATIONS IN THE SLEEPING AND AROUSED BRAIN
68 VONKROSIGK M,SCIENCE,vol 0261,page 0361,1993,cites= 33,CELLULAR MECHANISMS OF A SYNCHRONIZED OSCILLATION IN THE THALAMUS
69 JAHNSEN H,J PHYSL LON,vol 0349,page 0205,1984,cites= 39,ELECTROPHYSIOLOGICAL PROPERTIES OF GUINEA-PIG THALAMIC NEURONS - AN INVITRO STUDY
70 MCCORMICK DA,J PHYSL LON,vol 0431,page 0291,1990,cites= 48,PROPERTIES OF A HYPERPOLARIZATION-
ACTIVATED CATION CURRENT AND ITS ROLE IN RHYTHMIC OSCILLATION IN THALAMIC RELAY NEURONS
71 MAYER ML,J PHYSL LON,vol 0340,page 0019,1983,cites= 21,A VOLTAGE-CLAMP ANALYSIS OF INWARD%ANOMALOUS< RECTIFICATION IN MOUSE SPINAL SENSORY GANGLION NEURONS
72 HALLIWELL JV,BRAIN RES,vol 0250,page 0071,1982,cites= 45,VOLTAGE-CLAMP ANALYSIS OF MUSCARINIC EXCITATION IN HIPPOCAMPAL-NEURONS
73 LANCASTER B,J NEURPHYSL,vol 0055,page 1268,1986,cites= 31,CALCIUM-DEPENDENT CURRENT GENERATING THE AFTERHYPERPOLARIZATION OF HIPPOCAMPAL-NEURONS
74 MADISON DV,J PHYSL LON,vol 0354,page 0319,1984,cites= 27,CONTROL OF THE REPETITIVE DISCHARGE OF RAT CA1 PYRAMIDAL NEURONS INVITRO
75 STUART GJ,NATURE,vol 0367,page 0069,1994,cites= 72,ACTIVE PROPAGATION OF SOMATIC ACTION-
POTENTIALS INTO NEOCORTICAL PYRAMIDAL CELL DENDRITES
76 EDWARDS FA,PFLUG ARCH,vol 0414,page 0600,1989,cites= 97,A THIN SLICE PREPARATION FOR PATCH CLAMP RECORDINGS FROM NEURONS OF THE MAMMALIAN CENTRAL NERVOUS-SYSTEM

77 BLANTON MG,J NEUROSC M,vol 0030,page 0203,1989,cites= 96,WHOLE CELL RECORDING FROM NEURONS IN SLICES OF REPTILIAN AND MAMMALIAN CEREBRAL-CORTEX

SYNAPTIC TRANSMISSION

78 EDWARDS FA,J PHYSL LON,vol 0430,page 0213,1990,cites= 37,QUANTAL ANALYSIS OF INHIBITORY SYNAPTIC TRANSMISSION IN THE DENTATE GYRUS OF RAT HIPPOCAMPAL SLICES - A PATCH-CLAMP STUDY
79 BEKKERS JM,P NAS US,vol 0087,page 5359,1990,cites= 29,ORIGIN OF VARIABILITY IN QUANTAL SIZE IN CULTURED HIPPOCAMPAL-NEURONS AND HIPPOCAMPAL SLICES
80 FABER DS,SCIENCE,vol 0258,page 1494,1992,cites= 23,INTRINSIC QUANTAL VARIABILITY DUE TO STOCHASTIC PROPERTIES OF RECEPTOR-TRANSMITTER INTERACTIONS
81 BARBOUR B,NEURON,vol 0012,page 1331,1994,cites= 27,PROLONGED PRESENCE OF GLUTAMATE DURING EXCITATORY SYNAPTIC TRANSMISSION TO CEREBELLAR PURKINJE-CELLS
82 TRUSSELL LO,NEURON,vol 0010,page 1185,1993,cites= 30,DESENSITIZATION OF AMPA RECEPTORS UPON MULTIQUANTAL NEUROTRANSMITTER RELEASE
83 TRUSSELL LO,NEURON,vol 0003,page 0209,1989,cites= 29,GLUTAMATE RECEPTOR DESENSITIZATION AND ITS ROLE IN SYNAPTIC TRANSMISSION
84 YAMADA KA,J NEUROSC,vol 0013,page 3904,1993,cites= 43,BENZOTHIADIAZIDES INHIBIT RAPID GLUTAMATE-
RECEPTOR DESENSITIZATION AND ENHANCE GLUTAMATERGIC SYNAPTIC CURRENTS
85 PARTIN KM,NEURON,vol 0011,page 1069,1993,cites= 37,SELECTIVE MODULATION OF DESENSITIZATION AT AMPA VERSUS KAINATE RECEPTORS BY CYCLOTHIAZIDE AND CONCANAVALIN-A
86 HOLLMANN M,ANN R NEUR,vol 0017,page 0031,1994,cites= 233,CLONED GLUTAMATE RECEPTORS
87 CHOI DW,NEURON,vol 0001,page 0623,1988,cites= 259,GLUTAMATE NEUROTOXICITY AND DISEASES OF THE NERVOUS-SYSTEM

NITRIC-OXIDE MESSENGER

88 DAWSON VL,P NAS US,vol 0088,page 6368,1991,cites= 199,NITRIC-OXIDE MEDIATES GLUTAMATE NEUROTOXICITY IN PRIMARY CORTICAL CULTURES
89 GARTHWAITE J,TRENDS NEUR,vol 0014,page 0060,1991,cites= 206,GLUTAMATE, NITRIC-OXIDE AND CELL CELL SIGNALING IN THE NERVOUS-SYSTEM
90 BREDT DS,NEURON,vol 0008,page 0003,1992,cites= 171,NITRIC-OXIDE, A NOVEL NEURONAL MESSENGER
91 DAWSON TM,P NAS US,vol 0088,page 7797,1991,cites= 197,NITRIC-OXIDE SYNTHASE AND NEURONAL NADPH DIAPHORASE ARE IDENTICAL IN BRAIN AND PERIPHERAL-TISSUES
92 HOPE BT,P NAS US,vol 0088,page 2811,1991,cites= 198,NEURONAL NADPH DIAPHORASE IS A NITRIC-OXIDE SYNTHASE
93 BREDT DS,NATURE,vol 0351,page 0714,1991,cites= 183,CLONED AND EXPRESSED NITRIC-OXIDE SYNTHASE STRUCTURALLY RESEMBLES CYTOCHROME-P-450 REDUCTASE
94 LAMAS S,P NAS US,vol 0089,page 6348,1992,cites= 93,ENDOTHELIAL NITRIC-OXIDE SYNTHASE - MOLECULAR-
CLONING AND CHARACTERIZATION OF A DISTINCT CONSTITUTIVE ENZYME ISOFORM
95 JANSSENS SP,J BIOL CHEM,vol 0267,page 4519,1992,cites= 65,CLONING AND EXPRESSION OF A CDNA-ENCODING HUMAN ENDOTHELIUM-DERIVED RE-

LAXING FACTOR NITRIC-OXIDE SYNTHASE
96 NAKANE M,FEBS LETTER,vol 0316,page 0175,1993,cites= 53,CLONED HUMAN BRAIN NITRIC-OXIDE SYNTHASE IS HIGHLY EXPRESSED IN SKELETAL-MUSCLE
97 GELLER DA,P NAS US,vol 0090,page 3491,1993,cites= 118,MOLECULAR-CLONING AND EXPRESSION OF INDUCIBLE NITRIC-OXIDE SYNTHASE FROM HUMAN HEPATOCYTES
98 CHARTRAIN NA,J BIOL CHEM,vol 0269,page 6765,1994,cites= 49,MOLECULAR-CLONING, STRUCTURE, AND CHROMOSOMAL LOCALIZATION OF THE HUMAN INDUCIBLE NITRIC-OXIDE SYNTHASE GENE
99 LOWENSTEIN CJ,P NAS US,vol 0090,page 9730,1993,cites= 70,MACROPHAGE NITRIC-OXIDE SYNTHASE GENE - 2 UPSTREAM REGIONS MEDIATE INDUCTION BY INTERFERON-GAMMA AND LIPOPOLYSACCHARIDE
100 XIE QW,J EXP MED,vol 0177,page 1779,1993,cites= 81,PROMOTER OF THE MOUSE GENE ENCODING CALCIUM-
INDEPENDENT NITRIC-OXIDE SYNTHASE CONFERS INDUCIBILITY BY INTERFERON-GAMMA AND BACTERIAL LIPOPOLYSACCHARIDE
101 XIE QW,J BIOL CHEM,vol 0269,page 4705,1994,cites= 82,ROLE OF TRANSCRIPTION FACTOR NF-KAPPA-B/REL IN INDUCTION OF NITRIC-OXIDE SYNTHASE

BIOMEDICINE

NUCLEAR FACTOR KAPPA-B

102 SCHRECK R,J EXP MED,vol 0175,page 1181,1992,cites= 101,DITHIOCARBAMATES AS POTENT INHIBITORS OF NUCLEAR FACTOR KAPPA-B ACTIVATION IN INTACT-CELLS
103 SCHRECK R,EMBO J,vol 0010,page 2247,1991,cites= 256,REACTIVE OXYGEN INTERMEDIATES AS APPARENTLY WIDELY USED MESSENGERS IN THE ACTIVATION OF THE NF-KAPPA-B TRANSCRIPTION FACTOR AND HIV-1
104 BAEUERLE PA,ANN R IMMUN,vol 0012,page 0141,1994,cites= 221,FUNCTION AMD ACTIVATION OF NF-KAPPA-
B IN THE IMMUNE-SYSTEM

TUMOR NECROSIS FACTOR

105 OSBORN L,P NAS US,vol 0086,page 2336,1989,cites= 98,TUMOR NECROSIS FACTOR-ALPHA AND INTERLEUKIN-1 STIMULATE THE HUMAN IMMUNODEFICIENCY VIRUS ENHANCER BY ACTIVATION OF THE NUCLEAR FACTOR KAPPA-B
106 POLI G,P NAS US,vol 0087,page 0782,1990,cites= 36,TUMOR NECROSIS FACTOR-ALPHA FUNCTIONS IN AN AUTOCRINE MANNER IN THE INDUCTION OF HUMAN IMMUNODEFICIENCY VIRUS EXPRESSION
107 POLI G,J EXP MED,vol 0172,page 0151,1990,cites= 37,INTERLEUKIN-6 INDUCES HUMAN-IMMUNODEFICIENCY-
VIRUS EXPRESSION IN INFECTED MONOCYTIC CELLS ALONE AND IN SYNERGY WITH TUMOR NECROSIS FACTOR-ALPHA BY TRANSCRIPTIONAL AND POSTTRANSCRIPTIONAL MECHANISMS
108 BREEN EC,J IMMUNOL,vol 0144,page 0480,1990,cites= 39,INFECTION WITH HIV IS ASSOCIATED WITH ELEVATED IL-6 LEVELS AND PRODUCTION

INTERLEUKIN

109 CLERICI M,SCIENCE,vol 0262,page 1721,1993,cites= 69,RESTORATION OF HIV-SPECIFIC CELL-MEDIATED IMMUNE-RESPONSES BY INTERLEUKIN-12 IN-VITRO
110 CHEHIMI J,J EXP MED,vol 0179,page 1361,1994,cites= 57,IMPAIRED INTERLEUKIN-12 PRODUCTION IN HUMAN IMMUNODEFICIENCY VIRUS-INFECTED PATIENTS
111 CLERICI M,J CLIN INV,vol 0091,page 0759,1993,cites= 56,CHANGES IN

INTERLEUKIN-2 AND INTERLEUKIN-4 PRODUCTION IN ASYMPTOMATIC, HUMAN IMMUNODEFICIENCY VIRUS-SEROPOSITIVE INDIVIDUALS
112 GRAZIOSI C,SCIENCE,vol 0265,page 0248,1994,cites= 64,LACK OF EVIDENCE FOR THE DICHOTOMY OF T%H<1 AND T%H<2 PREDOMINANCE IN HIV-INFECTED INDIVIDUALS
113 MAGGI E,SCIENCE,vol 0265,page 0244,1994,cites= 67,ABILITY OF HIV TO PROMOTE A T%H<1 TO T%H<0 SHIFT AND TO REPLICATE PREFERENTIALLY IN T%H<2 AND T%H<0 CELLS
114 CLERICI M,IMMUNOL TOD,vol 0014,page 0107,1993,cites= 159,A T%H<1-*T%H<2 SWITCH IS A CRITICAL STEP IN THE ETIOLOGY OF HIV-INFECTION

DRUG RESISTANT HIV

115 PANTALEO G,NATURE,vol 0362,page 0355,1993,cites= 204,HIV-INFECTION IS ACTIVE AND PROGRESSIVE IN LYMPHOID-TISSUE DURING THE CLINICALLY LATENT STAGE OF DISEASE
116 HO DD,NATURE,vol 0373,page 0123,1995,cites= 421,RAPID TURNOVER OF PLASMA VIRIONS AND CD4 LYMPHOCYTES IN HIV-1 INFECTION
117 MELLORS JW,ANN INT MED,vol 0122,page 0573,1995,cites= 101,QUANTITATION OF HIV-1 RNA IN PLASMA PREDICTS OUTCOME AFTER SEROCONVERSION
118 OBRIEN WA,N ENG J MED,vol 0334,page 0426,1996,cites= 49,CHANGES IN PLASMA HIV-1 RNA AND CD4+ LYMPHOCYTE COUNTS AND THE RISK OF PROGRESSION TO AIDS
119 ERON JJ,N ENG J MED,vol 0333,page 1662,1995,cites= 49,TREATMENT WITH LAMIVUDINE, ZIDOVUDINE, OR BOTH IN HIV-POSITIVE PATIENTS WITH 200 TO 500 CD4+ CELLS PER CUBIC MILLIMETER
120 SCHUURMAN R,J INFEC DIS,vol 0171,page 1411,1995,cites= 38,RAPID CHANGES IN HUMAN-
IMMUNODEFICIENCY-VIRUS TYPE-1 RNA LOAD AND APPEARANCE OF DRUG-RESISTANT VIRUS POPULATIONS IN PERSONS TREATED WITH LAMIVUDINE %3TC<
121 TISDALE M,P NAS US,vol 0090,page 5653,1993,cites= 64,RAPID INVITRO SELECTION OF HUMAN-
IMMUNODEFICIENCY-VIRUS TYPE-1 RESISTANT TO 3$-THIACYTIDINE INHIBITORS DUE TO A MUTATION IN THE YMDD REGION OF REVERSE-TRANSCRIPTASE
122 STCLAIR MH,SCIENCE,vol 0253,page 1557,1991,cites= 70,RESISTANCE TO DDI AND SENSITIVITY TO AZT INDUCED BY A MUTATION IN HIV-1 REVERSE-TRANSCRIPTASE
123 LARDER BA,SCIENCE,vol 0246,page 1155,1989,cites= 87,MULTIPLE MUTATIONS IN HIV-1 REVERSE-
TRANSCRIPTASE CONFER HIGH-LEVEL RESISTANCE TO ZIDOVUDINE %AZT<
124 KOHLSTAEDT LA,SCIENCE,vol 0256,page 1783,1992,cites= 141,CRYSTAL-STRUCTURE AT 3.5 ANGSTROM RESOLUTION OF HIV-1 REVERSE-TRANSCRIPTASE COMPLEXED WITH AN INHIBITOR
125 SAWAYA MR,SCIENCE,vol 0264,page 1930,1994,cites= 41,CRYSTAL-STRUCTURE OF RAT DNA-POLYMERASE-
BETA - EVIDENCE FOR A COMMON POLYMERASE MECHANISM
126 OLLIS DL,NATURE,vol 0313,page 0762,1985,cites= 52,STRUCTURE OF LARGE FRAGMENT OF ESCHERICHIA-
COLI DNA-POLYMERASE-I COMPLEXED WITH DTMP
127 BEESE LS,SCIENCE,vol 0260,page 0352,1993,cites= 35,STRUCTURE OF DNA-POLYMERASE-I KLENOW FRAGMENT BOUND TO DUPLEX DNA

128 JOYCE CM,ANN R BIOCH,vol 0063,page 0777,1994,cites= 46,FUNCTION AND STRUCTURE RELATIONSHIPS IN DNA-POLYMERASES

PROTEINS

PROTEIN STRUCTURE

129 BEESE LS,EMBO J,vol 0010,page 0025,1991,cites= 59,STRUCTURAL BASIS FOR THE 3-5 EXONUCLEASE ACTIVITY OF ESCHERICHIA-COLI DNA-POLYMERASE-I - A 2 METAL-ION MECHANISM

130 KRAULIS PJ,J APPL CRYS,vol 0024,page 0946,1991,cites= 807,MOLSCRIPT - A PROGRAM TO PRODUCE BOTH DETAILED AND SCHEMATIC PLOTS OF PROTEIN STRUCTURES

131 JONES TA,ACT CRYST A,vol 0047,page 0110,1991,cites= 391,IMPROVED METHODS FOR BUILDING PROTEIN MODELS IN ELECTRON-DENSITY MAPS AND THE LOCATION OF ERRORS IN THESE MODELS

132 TRONRUD DE,ACT CRYST A,vol 0043,page 0489,1987,cites= 79,AN EFFICIENT GENERAL-PURPOSE LEAST-
SQUARES REFINEMENT PROGRAM FOR MACROMOLECULAR STRUCTURES

133 JONES TA,METH ENZYM,vol 0115,page 0157,1985,cites= 95,INTERACTIVE COMPUTER-GRAPHICS - FRODO

134 KABSCH W,J APPL CRYS,vol 0021,page 0067,1988,cites= 39,AUTOMATIC-INDEXING OF ROTATION DIFFRACTION PATTERNS

135 KABSCH W,J APPL CRYS,vol 0021,page 0916,1988,cites= 145,EVALUATION OF SINGLE-CRYSTAL X-RAY-
DIFFRACTION DATA FROM A POSITION-SENSITIVE DETECTOR

CHEMISTRY

136 SHELDRICK GM,ACT CRYST A,vol 0046,page 0467,1990,cites= 970,PHASE ANNEALING IN SHELX-90 - DIRECT METHODS FOR LARGER STRUCTURES

METALLOCENE CATALYSTS

137 BRINTZINGER HH,ANGEW CHEM,vol 0034,page 1143,1995,cites= 77,STEREOSPECIFIC OLEFIN POLYMERIZATION WITH CHIRAL METALLOCENE CATALYSTS

138 MOHRING PC,J ORGMET CH,vol 0479,page 0001,1994,cites= 45,HOMOGENEOUS GROUP-4 METALLOCENE ZIEGLER-NATTA CATALYSTS - THE INFLUENCE OF CYCLOPENTADIENYL-RING SUBSTITUENTS

139 MARKS TJ,ACC CHEM RE,vol 0025,page 0057,1992,cites= 22,SURFACE-BOUND METAL HYDROCARBYLS - ORGANOMETALLIC CONNECTIONS BETWEEN HETEROGENEOUS AND HOMOGENEOUS CATALYSIS

140 JORDAN RF,ADV ORGMET,vol 0032,page 0325,1991,cites= 43,CHEMISTRY OF CATIONIC DICYCLOPENTADIENYL GROUP-4 METAL ALKYL COMPLEXES

KHAND REACTION

141 BUCHWALD SL,CHEM REV,vol 0088,page 1047,1988,cites= 35,GROUP-4 METAL-COMPLEXES OF BENZYNES, CYCLOALKYNES, ACYCLIC ALKYNES, AND ALKENES

142 URABE H,TETRAHEDR L,vol 0036,page 4261,1995,cites= 14,SYNTHETIC APPLICATION OF TITANABICYCLES GENERATED FROM 1,6-DIENES OR 1,7-DIENES, ENYNES, AND DIYNES AND %ETA%2<-PROPENE<TI%O-I-
PR<%2<

143 BERK SC,J AM CHEM S,vol 0116,page 8593,1994,cites= 17,DEVELOPMENT OF A TITANOCENE-CATALYZED ENYNE CYCLIZATION ISOCYANIDE INSERTION REACTION

144 JEONG N,J AM CHEM S,vol 0116,page 3159,1994,cites= 16,CATALYTIC VERSION OF THE INTRAMOLECULAR PAUSON-KHAND REACTION

145 PAUSON PL,TETRAHEDRON,vol 0041,page 5855,1985,cites= 23,THE KHAND REACTION - A CONVENIENT AND GENERAL-ROUTE TO A WIDE-RANGE OF CYCLOPENTENONE DERIVATIVES
146 SHAMBAYATI S,TETRAHEDR L,vol 0031,page 5289,1990,cites= 14,N-OXIDE PROMOTED PAUSON-KHAND CYCLIZATIONS AT ROOM-TEMPERATURE

PALLADIUM CATALYSIS

147 OPPOLZER W,ANGEW CHEM,vol 0028,page 0038,1989,cites= 17,INTRAMOLECULAR, STOICHIOMETRIC %LI, MG, ZN< AND CATALYTIC %NI, PD, PT< METALLO-ENE REACTIONS IN ORGANIC-SYNTHESIS
148 TROST BM,ACC CHEM RE,vol 0023,page 0034,1990,cites= 21,PALLADIUM-CATALYZED CYCLOISOMERIZATIONS OF ENYNES AND RELATED REACTIONS
149 TROST BM,J AM CHEM S,vol 0116,page 4255,1994,cites= 18,PD-CATALYZED CYCLOISOMERIZATION TO 1,2-
DIALKYLIDENECYCLOALKANES .1.
150 TROST BM,J AM CHEM S,vol 0116,page 4268,1994,cites= 13,PD-CATALYZED CYCLOISOMERIZATION TO 1,2-
DIALKYLIDENECYCLOALKANES .2. ALTERNATIVE CATALYST SYSTEM
151 TROST BM,J AM CHEM S,vol 0115,page 9421,1993,cites= 16,PALLADIUM-CATALYZED CYCLIZATIONS OF POLYENYNES - A PALLADIUM ZIPPER
152 OVERMAN LE,J AM CHEM S,vol 0115,page 2042,1993,cites= 8,1ST TOTAL SYNTHESIS OF SCOPADULCIC ACID-B
153 GRIGG R,TETRAHEDR L,vol 0031,page 1343,1990,cites= 14,PALLADIUM-CATALYZED POLYCYCLISATION-
ANION CAPTURE PROCESSES
154 HECK RF,ORG REACT,vol 0027,page 0345,1982,cites= 64,PALLADIUM-CATALYZED VINYLATION OF ORGANIC HALIDES
155 STILLE JK,ANGEW CHEM,vol 0025,page 0508,1986,cites= 132,THE PALLADIUM-CATALYZED CROSS-COUPLING REACTIONS OF ORGANOTIN REAGENTS WITH ORGANIC ELECTROPHILES

ASYMMETRIC CATALYSIS

156 TSUJI J,TETRAHEDRON,vol 0042,page 4361,1986,cites= 25,NEW GENERAL SYNTHETIC METHODS INVOLVING PI-ALLYLPALLADIUM COMPLEXES AS INTERMEDIATES AND NEUTRAL REACTION CONDITIONS
157 CONSIGLIO G,CHEM REV,vol 0089,page 0257,1989,cites= 36, ENANTIOSELECTIVE HOMOGENEOUS CATALYSIS INVOLVING TRANSITION-METAL ALLYL INTERMEDIATES
158 BROWN JM,TETRAHEDRON,vol 0050,page 4493,1994,cites= 30,MECHANISTIC AND SYNTHETIC STUDIES IN CATALYTIC ALLYLIC ALKYLATION WITH PALLADIUM COMPLEXES OF 1-%2-DIPHENYLPHOSPHINO -1-NAPHTHYL<ISOQUINOLINE
159 SPRINZ J,TETRAHEDR L,vol 0035,page 1523,1994,cites= 35,CATALYSIS OF ALLYLIC SUBSTITUTIONS BY PD COMPLEXES OF OXAZOLINES CONTAINING AN ADDITIONAL P, S, OR SE CENTER - X-RAY CRYSTAL-
STRUCTURES AND SOLUTION STRUCTURES OF CHIRAL PI-ALLYL PALLADIUM COMPLEXES OF PHOSPHINOARYLOXAZOLINES
160 LEUTENEGGER U,TETRAHEDRON,vol 0048,page 2143,1992,cites= 27,5-AZA-SEMICORRINS - A NEW CLASS OF BIDENTATE NITROGEN LIGANDS FOR ENANTIOSELECTIVE CATALYSIS
161 PFALTZ A,ACC CHEM RE,vol 0026,page 0339,1993,cites= 51,CHIRAL SEMICORRINS AND RELATED NITROGEN-HETEROCYCLES AS LIGANDS IN ASYMMETRIC CATALYSIS
162 BOLM C,ANGEW CHEM,vol 0030,page 0542,1991,cites= 15,BIS%4,5-

DIHYDROOXAZOLYL< DERIVATIVES IN ASYMMETRIC CATALYSIS
163 HELMCHEN G,SYNLETT,vol ,page 0257,1991,cites= 16,C-2 SYMMETRICAL BIOXAZOLINES AND BITHIAZOLINES AS NEW CHIRAL LIGANDS FOR METAL-ION CATALYZED ASYMMETRIC SYNTHESES - ASYMMETRIC HYDROSILYLATION
164 LOWENTHAL RE,TETRAHEDR L,vol 0031,page 6005,1990,cites= 35,ASYMMETRIC CATALYTIC CYCLOPROPANATION OF OLEFINS - BIS-OXAZOLINE COPPER-COMPLEXES
165 EVANS DA,J AM CHEM S,vol 0113,page 0726,1991,cites= 44,BIS%OXAZOLINES< AS CHIRAL LIGANDS IN METAL-CATALYZED ASYMMETRIC REACTIONS - CATALYTIC, ASYMMETRIC CYCLOPROPANATION OF OLEFINS
166 LOWENTHAL RE,TETRAHEDR L,vol 0032,page 7373,1991,cites= 22,ASYMMETRIC COPPER-CATALYZED CYCLOPROPANATION OF TRISUBSTITUTED AND UNSYMMETRICAL CIS-1,2-DISUBSTITUTED OLEFINS - MODIFIED BIS-OXAZOLINE LIGANDS
167 EVANS DA,J AM CHEM S,vol 0115,page 5328,1993,cites= 35,BIS%OXAZOLINE< COPPER-COMPLEXES AS CHIRAL CATALYSTS FOR THE ENANTIOSELECTIVE AZIRIDINATION OF OLEFINS
168 LI Z,J AM CHEM S,vol 0115,page 5326,1993,cites= 29,ASYMMETRIC ALKENE AZIRIDINATION WITH READILY AVAILABLE CHIRAL DIIMINE-BASED CATALYSTS

PORPHYRINS

169 JACOBSEN EN,J AM CHEM S,vol 0113,page 7063,1991,cites= 46,HIGHLY ENANTIOSELECTIVE EPOXIDATION CATALYSTS DERIVED FROM 1,2-DIAMINOCYCLOHEXANE
170 PALUCKI M,J AM CHEM S,vol 0116,page 9333,1994,cites= 20,HIGHLY ENANTIOSELECTIVE, LOW-
TEMPERATURE EPOXIDATION OF STYRENE
171 GROVES JT,J AM CHEM S,vol 0105,page 5791,1983,cites= 22,CATALYTIC ASYMMETRIC EPOXIDATIONS WITH CHIRAL IRON PORPHYRINS
172 COLLMAN JP,SCIENCE,vol 0261,page 1404,1993,cites= 37,REGIOSELECTIVE AND ENANTIOSELECTIVE EPOXIDATION CATALYZED BY METALLOPORPHYRINS
173 MEUNIER B,CHEM REV,vol 0092,page 1411,1992,cites= 94, METALLOPORPHYRINS AS VERSATILE CATALYSTS FOR OXIDATION REACTIONS AND OXIDATIVE DNA CLEAVAGE
174 TRAYLOR PS,J CHEM S CH,vol ,page 0279,1984,cites= 23,STERICALLY PROTECTED HEMINS WITH ELECTRONEGATIVE SUBSTITUENTS - EFFICIENT CATALYSTS FOR HYDROXYLATION AND EPOXIDATION
175 TRAYLOR TG,INORG CHEM,vol 0026,page 1338,1987,cites= 23,PERHALOGENATED TETRAPHENYLHEMINS - STABLE CATALYSTS OF HIGH TURNOVER CATALYTIC HYDROXYLATIONS
176 TRAYLOR TG,J AM CHEM S,vol 0114,page 1308,1992,cites= 19,ALIPHATIC HYDROXYLATION CATALYZED BY IRON%III< PORPHYRINS
177 RENAUD JP,J CHEM S CH,vol ,page 0888,1985,cites= 10,A VERY EFFICIENT SYSTEM FOR ALKENE EPOXIDATION BY HYDROGEN-PEROXIDE - CATALYSIS BY MANGANESE-PORPHYRINS IN THE PRESENCE OF IMIDAZOLE
178 HOFFMANN P,B S CHIM FR,vol 0129,page 0085,1992,cites= 13,PREPARATION AND CATALYTIC ACTIVITIES OF THE MANGANESE AND IRON DERIVATIVES OF BR8TMP AND CL12TMP, 2 ROBUST PORPHYRIN LIGANDS OBTAINED BY HALOGENATION OF TETRAMESITYLPORPHYRIN
179 LINDSEY JS, J ORG CHEM,vol 0054, page 0828, 1989, cites= 44, INVESTIGA-

TION OF THE SYNTHESIS OF ORTHO-SUBSTITUTED TETRAPHENYLPORPHYRINS
180 LINDSEY JS,J ORG CHEM,vol 0052,page 0827,1987,cites= 42,ROTHEMUND AND ADLER-LONGO REACTIONS REVISITED - SYNTHESIS OF TETRAPHENYLPORPHYRINS UNDER EQUILIBRIUM CONDITIONS
181 LINDSEY JS,TETRAHEDRON,vol 0050,page 8941,1994,cites= 14,PORPHYRIN BUILDING-BLOCKS FOR MODULAR CONSTRUCTION OF BIOORGANIC MODEL SYSTEMS
182 SETH J,J AM CHEM S,vol 0116,page 0578,1994,cites= 24,INVESTIGATION OF ELECTRONIC COMMUNICATION IN MULTI-PORPHYRIN LIGHT-HARVESTING ARRAYS
183 WAGNER RW,J AM CHEM S,vol 0116,page 9759,1994,cites= 34,A MOLECULAR PHOTONIC WIRE
184 PRATHAPAN S,J AM CHEM S,vol 0115,page 7519,1993,cites= 25,BUILDING-BLOCK SYNTHESIS OF PORPHYRIN LIGHT-HARVESTING ARRAYS
185 SESSLER JL,J AM CHEM S,vol 0115,page 4618,1993,cites= 23,ELECTRONIC-ENERGY MIGRATION AND TRAPPING IN QUINONE-SUBSTITUTED, PHENYL-LINKED DIMERIC AND TRIMERIC PORPHYRINS
186 OSUKA A,J AM CHEM S,vol 0115,page 4577,1993,cites= 17,1,2-PHENYLENE-BRIDGED DIPORPHYRIN LINKED WITH PORPHYRIN MONOMER AND PYROMELLITIMIDE AS A MODEL FOR A PHOTOSYNTHETIC REACTION CENTER - SYNTHESIS AND PHOTOINDUCED CHARGE SEPARATION
187 JOHNSON DG,J AM CHEM S,vol 0115,page 5692,1993,cites= 15,PHOTOCHEMICAL ELECTRON-TRANSFER IN CHLOROPHYLL PORPHYRIN QUINONE TRIADS - THE ROLE OF THE PORPHYRIN-BRIDGING MOLECULE
188 GUST D,ACC CHEM RE,vol 0026,page 0198,1993,cites= 50,MOLECULAR MIMICRY OF PHOTOSYNTHETIC ENERGY AND ELECTRON-TRANSFER
189 WASIELEWSKI MR,CHEM REV,vol 0092,page 0435,1992,cites= 118,PHOTOINDUCED ELECTRON-TRANSFER IN SUPRAMOLECULAR SYSTEMS FOR ARTIFICIAL PHOTOSYNTHESIS

ECOLOGY

190 MCDERMOTT G,NATURE,vol 0374,page 0517,1995,cites= 123,CRYSTAL-STRUCTURE OF AN INTEGRAL MEMBRANE LIGHT-HARVESTING COMPLEX FROM PHOTOSYNTHETIC BACTERIA
191 KUHLBRANDT W,NATURE,vol 0367,page 0614,1994,cites= 93,ATOMIC MODEL OF PLANT LIGHT-HARVESTING COMPLEX BY ELECTRON CRYSTALLOGRAPHY
192 DEMMIGADAMS B,BIOC BIOP A,vol 1020,page 0001,1990,cites= 75,CAROTENOIDS AND PHOTOPROTECTION IN PLANTS - A ROLE FOR THE XANTHOPHYLL ZEAXANTHIN

ATMOSPHERIC C02

193 SCHREIBER U,PHOTOSYN R,vol 0025,page 0279,1990,cites= 20,O-2-DEPENDENT ELECTRON FLOW, MEMBRANE ENERGIZATION AND THE MECHANISM OF NONPHOTOCHEMICAL QUENCHING OF CHLOROPHYLL FLUORESCENCE
194 MIYAKE C,PLANT CEL P,vol 0033,page 0541,1992,cites= 20,THYLAKOID-BOUND ASCORBATE PEROXIDASE INSPINACH-CHLOROPLASTS AND PHOTOREDUCTION OF ITS PRIMARY OXIDATION-PRODUCT MONODEHYDROASCORBATE RADICALS IN THYLAKOIDS
195 MIYAKE C,PLANT CEL P,vol 0034,page 0881,1993,cites= 12,PURIFICATION AND MOLECULAR-PROPERTIES OF THE THYLAKOID-BOUND ASCORBATE PEROXIDASE IN SPINACH-CHLOROPLASTS
196 MITTLER R,PLANT PHYSL,vol 0097,page 0962,1991,cites= 12,PURIFICATION

AND CHARACTERIZATION OF PEA CYTOSOLIC ASCORBATE PEROXIDASE
197 NAKANO Y,PLANT CEL P,vol 0028,page 0131,1987,cites= 18,PURIFICATION OF ASCORBATE PEROXIDASE IN SPINACH-CHLOROPLASTS - ITS INACTIVATION IN ASCORBATE-DEPLETED MEDIUM AND REACTIVATION BY MONODEHYDROASCORBATE RADICAL
198 POLLE A,PLANT PHYSL,vol 0094,page 0312,1990,cites= 16,COMPOSITION AND PROPERTIES OF HYDROGEN-
PEROXIDE DECOMPOSING SYSTEMS IN EXTRACELLULAR AND TOTAL EXTRACTS FROM NEEDLES OF NORWAY SPRUCE %PICEA-ABIES L, KARST<
199 TAKAHAMA U,PLANT CEL P,vol 0033,page 0379,1992,cites= 11,REGULATION OF PEROXIDASE-DEPENDENT OXIDATION OF PHENOLICS IN THE APOPLAST OF SPINACH LEAVES BY ASCORBATE
200 LUWE MWF,PLANT PHYSL,vol 0101,page 0969,1993,cites= 19,ROLE OF ASCORBATE IN DETOXIFYING OZONE IN THE APOPLAST OF SPINACH %SPINACIA-OLERACEA L< LEAVES
201 KANGASJARVI J,PL CELL ENV,vol 0017,page 0783,1994,cites= 24,PLANT DEFENSE SYSTEMS INDUCED BY OZONE
202 DARRALL NM,PLANT CELL,vol 0012,page 0001,1989,cites= 32,THE EFFECT OF AIR-POLLUTANTS ON PHYSIOLOGICAL PROCESSES IN PLANTS
203 MATYSSEK R,TREES,vol 0005,page 0005,1991,cites= 13,IMPAIRMENT OF GAS-EXCHANGE AND STRUCTURE IN BIRCH LEAVES %BETULA-PENDULA< CAUSED BY LOW OZONE CONCENTRATIONS
204 REICH PB,PLANT PHYSL,vol 0073,page 0291,1983,cites= 16,EFFECTS OF LOW CONCENTRATIONS OF O-3 ON NET PHOTOSYNTHESIS, DARK RESPIRATION, AND CHLOROPHYLL CONTENTS IN AGING HYBRID POPLAR LEAVES
205 REICH PB,SCIENCE,vol 0230,page 0566,1985,cites= 14,AMBIENT LEVELS OF OZONE REDUCE NET PHOTOSYNTHESIS IN TREE AND CROP SPECIES
206 NIE GY,PL CELL ENV,vol 0016,page 0643,1993,cites= 8,EFFECTS OF OZONE ON THE PHOTOSYNTHETIC APPARATUS AND LEAF PROTEINS DURING LEAF DEVELOPMENT IN WHEAT
207 FARAGE PK,PLANT PHYSL,vol 0095,page 0529,1991,cites= 15,THE SEQUENCE OF CHANGE WITHIN THE PHOTOSYNTHETIC APPARATUS OF WHEAT FOLLOWING SHORT-TERM EXPOSURE TO OZONE
208 BARNES JD,NEW PHYTOL,vol 0121,page 0403,1992,cites= 11,THE INFLUENCE OF CO2 AND O-3, SINGLY AND IN COMBINATION, ON GAS-EXCHANGE, GROWTH AND NUTRIENT STATUS OF RADISH %RAPHANUS-SATIVUS L<
209 ALLEN LH,J ENVIR Q,vol 0019,page 0015,1990,cites= 18,PLANT-RESPONSES TO RISING CARBON-DIOXIDE AND POTENTIAL INTERACTIONS WITH AIR-POLLUTANTS
210 CURE JD,AGR FOR MET,vol 0038,page 0127,1986,cites= 50,CROP RESPONSES TO CARBON-DIOXIDE DOUBLING - A LITERATURE SURVEY
211 STITT M,PL CELL ENV,vol 0014,page 0741,1991,cites= 65,RISING CO2 LEVELS AND THEIR POTENTIAL SIGNIFICANCE FOR CARBON FLOW IN PHOTOSYNTHETIC CELLS
212 SAGE RF,PLANT PHYSL,vol 0089,page 0590,1989,cites= 53,ACCLIMATION OF PHOTOSYNTHESIS TO ELEVATED CO2 IN 5 C-3 SPECIES
213 TISSUE DT,PL CELL ENV,vol 0016,page 0859,1993,cites= 34,LONG-TERM EFFECTS OF ELEVATED CO2 AND NUTRIENTS ON PHOTOSYNTHESIS AND RUBISCO IN LOBLOLLY-PINE SEEDLINGS
214 GUNDERSON CA,PHOTOSYN R,vol 0039,page 0369,1994,cites= 25,PHOTOSYNTHETIC ACCLIMATION IN TREES TO RISING ATMOSPHERIC CO2 - A BROADER PERSPECTIVE

215 CEULEMANS R,NEW PHYTOL,vol 0127,page 0425,1994,cites= 37,TANSLEY REVIEW NO-71 - EFFECTS OF ELEVATED ATMOSPHERIC CO2 ON WOODY-PLANTS
216 EAMUS D,ADV ECOL R,vol 0019,page 0001,1989,cites= 60,THE DIRECT EFFECTS OF INCREASE IN THE GLOBAL ATMOSPHERIC CO2 CONCENTRATION ON NATURAL AND COMMERCIAL TEMPERATE TREES AND FORESTS

MODELS OF BIOSPHERE

217 TANS PP,SCIENCE,vol 0247,page 1431,1990,cites= 91,OBSERVATIONAL CONSTRAINTS ON THE GLOBAL ATMOSPHERIC CO2 BUDGET
218 ENTING IG,TELLUS B,vol 0043,page 0156,1991,cites= 19,LATITUDINAL DISTRIBUTION OF SOURCES AND SINKS OF CO2 - RESULTS OF AN INVERSION STUDY
219 DENNING AS,NATURE,vol 0376,page 0240,1995,cites= 18,LATITUDINAL GRADIENT OF ATMOSPHERIC CO2 DUE TO SEASONAL EXCHANGE WITH LAND BIOTA
220 FUNG IY,J GEO RES-A,vol 0092,page 2999,1987,cites= 26,APPLICATION OF ADVANCED VERY HIGH-
RESOLUTION RADIOMETER VEGETATION INDEX TO STUDY ATMOSPHERE-BIOSPHERE EXCHANGE OF CO2
221 POTTER CS,GLOBAL BIOG,vol 0007,page 0811,1993,cites= 39,TERRESTRIAL ECOSYSTEM PRODUCTION - A PROCESS MODEL-BASED ON GLOBAL SATELLITE AND SURFACE DATA
222 PARTON WJ,GLOBAL BIOG,vol 0007,page 0785,1993,cites= 24,OBSERVATIONS AND MODELING OF BIOMASS AND SOIL ORGANIC-MATTER DYNAMICS FOR THE GRASSLAND BIOME WORLDWIDE
223 DUCOUDRE NI,J CLIMATE,vol 0006,page 0248,1993,cites= 12,SECHIBA, A NEW SET OF PARAMETERIZATIONS OF THE HYDROLOGIC EXCHANGES AT THE LAND ATMOSPHERE INTERFACE WITHIN THE LMD ATMOSPHERIC GENERAL-CIRCULATION MODEL
224 VERSEGHY DL,INT J CLIM,vol 0013,page 0347,1993,cites= 12,CLASS - A CANADIAN LAND-SURFACE SCHEME FOR GCMS .2. VEGETATION MODEL AND COUPLED RUNS
225 VERSEGHY DL,INT J CLIM,vol 0011,page 0111,1991,cites= 15,CLASS-A CANADIAN LAND SURFACE SCHEME FOR GCMS .1. SOIL MODEL
226 NOILHAN J,M WEATH REV,vol 0117,page 0536,1989,cites= 34,A SIMPLE PARAMETERIZATION OF LAND SURFACE PROCESSES FOR METEOROLOGICAL MODELS
227 SELLERS PJ,J ATMOS SCI,vol 0043,page 0505,1986,cites= 75,A SIMPLE BIOSPHERE MODEL %SIB< FOR USE WITHIN GENERAL-CIRCULATION MODELS

GEOSCIENCE

CLIMATE CYCLES

228 FOLEY JA,NATURE,vol 0371,page 0052,1994,cites= 18,FEEDBACKS BETWEEN CLIMATE AND BOREAL FORESTS DURING THE HOLOCENE EPOCH
229 KUTZBACH JE,J ATMOS SCI,vol 0043,page 1726,1986,cites= 37,THE INFLUENCE OF CHANGING ORBITAL PARAMETERS AND SURFACE BOUNDARY-CONDITIONS ON CLIMATE SIMULATIONS FOR THE PAST 18000 YEARS
230 PRELL WL,J GEO RES-A,vol 0092,page 8411,1987,cites= 23,MONSOON VARIABILITY OVER THE PAST 150,000 YEARS
231 CHAPPELLAZ J,NATURE,vol 0345,page 0127,1990,cites= 16,ICE-CORE RECORD OF ATMOSPHERIC METHANE OVER THE PAST 160,000 YEARS
232 SOWERS T,SCIENCE,vol 0269,page 0210,1995,cites= 8,CLIMATE RECORDS COVERING THE LAST DEGLACIATION
233 JOUZEL J,NATURE,vol 0329,page 0403,1987,cites= 13,VOSTOK ICE CORE - A

CONTINUOUS ISOTOPE TEMPERATURE RECORD OVER THE LAST CLIMATIC CYCLE %160,000 YEARS<
234 BROECKER WS,GEOCH COS A,vol 0053,page 2465,1989,cites= 39,THE ROLE OF OCEAN-ATMOSPHERE REORGANIZATIONS IN GLACIAL CYCLES
235 LEHMAN SJ,NATURE,vol 0356,page 0757,1992,cites= 35,SUDDEN CHANGES IN NORTH-ATLANTIC CIRCULATION DURING THE LAST DEGLACIATION
236 BOND G,NATURE,vol 0365,page 0143,1993,cites= 59,CORRELATIONS BETWEEN CLIMATE RECORDS FROM NORTH-ATLANTIC SEDIMENTS AND GREENLAND ICE
237 BOND G,NATURE,vol 0360,page 0245,1992,cites= 44,EVIDENCE FOR MASSIVE DISCHARGES OF ICEBERGS INTO THE NORTH-ATLANTIC OCEAN DURING THE LAST GLACIAL PERIOD

EARTH'S GEOID

238 MANGERUD J,QUAT SCI R,vol 0011,page 0633,1992,cites= 15,THE LAST INTERGLACIAL GLACIAL PERIOD ON SPITSBERGEN, SVALBARD
239 MANGERUD J,QUATERN RES,vol 0038,page 0001,1992,cites= 15,THE LAST GLACIAL MAXIMUM ON SPITSBERGEN, SVALBARD
240 ELVERHOI A,QUAT SCI R,vol 0012,page 0863,1993,cites= 16,THE BARENTS SEA-ICE SHEET - A MODEL OF ITS GROWTH AND DECAY DURING THE LAST ICE MAXIMUM
241 LAMBECK K,QUAT SCI R,vol 0014,page 0001,1995,cites= 9,CONSTRAINTS ON THE LATE WEICHSELIAN ICE-
SHEET OVER THE BARENTS SEA FROM OBSERVATIONS OF RAISED SHORELINES
242 LAMBECK K,GEOPHYS J I,vol 0103,page 0451,1990,cites= 11,HOLOCENE GLACIAL REBOUND AND SEA-LEVEL CHANGE IN NW EUROPE
243 MITROVICA JX,J GEO R-SOL,vol 0096,page 0053,1991,cites= 13,ON POSTGLACIAL GEOID SUBSIDENCE OVER THE EQUATORIAL OCEANS
244 MITROVICA JX,GEOPHYS J I,vol 0114,page 0045,1993,cites= 15,THE INFERENCE OF MANTLE VISCOSITY FROM AN INVERSION OF THE FENNOSCANDIAN RELAXATION SPECTRUM
245 PELTIER R,ADV GEOPHYS,vol 0024,page 0001,1982,cites= 10,DYNAMICS OF THE ICE-AGE EARTH
246 RICHARDS MA,J GEOPH RES,vol 0089,page 5987,1984,cites= 18,GEOID ANOMALIES IN A DYNAMIC EARTH
247 FORTE AM,J GEO R-SOL,vol 0096,page 0131,1991,cites= 15,VISCOUS-FLOW MODELS OF GLOBAL GEOPHYSICAL OBSERVABLES .1. FORWARD PROBLEMS
248 HAGER BH,PHI T ROY A,vol 0328,page 0309,1989,cites= 18,LONG-WAVELENGTH VARIATIONS IN EARTHS GEOID - PHYSICAL MODELS AND DYNAMICAL IMPLICATIONS
249 TACKLEY PJ,NATURE,vol 0361,page 0699,1993,cites= 29,EFFECTS OF AN ENDOTHERMIC PHASE-TRANSITION AT 670 KM DEPTH IN A SPHERICAL MODEL OF CONVECTION IN THE EARTHS MANTLE

EARTHS MANTLE

250 SU WJ,J GEO R-SOL,vol 0099,page 6945,1994,cites= 44,DEGREE-12 MODEL OF SHEAR VELOCITY HETEROGENEITY IN THE MANTLE
251 GRAND SP,J GEO R-SOL,vol 0099,page 1591,1994,cites= 32,MANTLE SHEAR STRUCTURE BENEATH THEAMERICA AND SURROUNDING OCEANS
252 VANDERHILST R,NATURE,vol 0374,page 0154,1995,cites= 19,COMPLEX MORPHOLOGY OF SUBDUCTED LITHOSPHERE IN THE MANTLE BENEATH THE TONGA TRENCH
253 FUKAO Y,J GEO R-SOL,vol 0097,page 4809,1992,cites= 19,SUBDUCTING SLABS

STAGNANT IN THE MANTLE TRANSITION ZONE
254 ITO E,J GEO R-S E,vol 0094,page 0637,1989,cites= 28,POSTSPINEL TRANSFORMATIONS IN THE SYSTEM MG2SIO4-FE2SIO4 AND SOME GEOPHYSICAL IMPLICATIONS
255 KATSURA T,J GEO R-S E,vol 0094,page 5663,1989,cites= 26,THE SYSTEM MG2SIO4-FE2SIO4 AT HIGH-
PRESSURES AND TEMPERATURES - PRECISE DETERMINATION OF STABILITIES OF OLIVINE, MODIFIED SPINEL, AND SPINEL

SEISMIC VELOCITIES

256 ITA J,J GEO R-SOL,vol 0097,page 6849,1992,cites= 15,PETROLOGY, ELASTICITY, AND COMPOSITION OF THE MANTLE TRANSITION ZONE
257 DUFFY TS,J GEO R-S E,vol 0094,page 1895,1989,cites= 14,SEISMIC VELOCITIES IN MANTLE MINERALS AND THE MINERALOGY OF THE UPPER MANTLE
258 ZAUG JM,SCIENCE,vol 0260,page 1487,1993,cites= 11,SOUND VELOCITIES IN OLIVINE AT EARTH MANTLE PRESSURES
259 WANG YB,PHYS E PLAN,vol 0083,page 0013,1994,cites= 18,P-V-T EQUATION OF STATE OF %MG,FE<SIO3 PEROVSKITE - CONSTRAINTS ON COMPOSITION OF THE LOWER MANTLE
260 MAO HK,J GEO R-S E,vol 0096,page 8069,1991,cites= 19,EFFECT OF PRESSURE, TEMPERATURE, AND COMPOSITION ON LATTICE-PARAMETERS AND DENSITY OF %FE,MG<SIO3-PEROVSKITES TO 30 GPA
261 ANDERSON OL,J APPL PHYS,vol 0065,page 1534,1989,cites= 13,ANHARMONICITY AND THE EQUATION OF STATE FOR GOLD
262 VINET P,PHYS REV B,vol 0035,page 1945,1987,cites= 10,TEMPERATURE EFFECTS ON THE UNIVERSAL EQUATION OF STATE OF SOLIDS

EQUATION OF STATE FOR SOLIDS

263 VINET P,J PHYS C,vol 0019,page 0467,1986,cites= 19,A UNIVERSAL EQUATION OF STATE FOR SOLIDS
264 LIU J,PHYS REV L,vol 0072,page 4105,1994,cites= 9,RAMAN MODES OF 6H POLYTYPE OF SILICON-CARBIDE TO ULTRAHIGH PRESSURES - A COMPARISON WITH SILICON AND DIAMOND
265 GIANNOZZI P,PHYS REV B,vol 0043,page 7231,1991,cites= 40,ABINITIO CALCULATION OF PHONON DISPERSIONS IN SEMICONDUCTORS
266 BARONI S,PHYS REV L,vol 0058,page 1861,1987,cites= 28,GREEN-FUNCTION APPROACH TO LINEAR RESPONSE IN SOLIDS
267 TROULLIER N,PHYS REV B,vol 0043,page 1993,1991,cites= 126,EFFICIENT PSEUDOPOTENTIALS FOR PLANE-
WAVE CALCULATIONS
268 KLEINMAN L,PHYS REV L,vol 0048,page 1425,1982,cites= 139,EFFICACIOUS FORM FOR MODEL PSEUDOPOTENTIALS

AB-INITIO MOLECULAR DYNAMICS

269 PAYNE MC,REV M PHYS,vol 0064,page 1045,1992,cites= 133,ITERATIVE MINIMIZATION TECHNIQUES FOR ABINITIO TOTAL-ENERGY CALCULATIONS - MOLECULAR-DYNAMICS AND CONJUGATE GRADIENTS
270 LIN JS,PHYS REV B,vol 0047,page 4174,1993,cites= 22,OPTIMIZED AND TRANSFERABLE NONLOCAL SEPARABLE ABINITIO PSEUDOPOTENTIALS
271 KINGSMITH RD,PHYS REV B,vol 0044,page 3063,1991,cites= 17,REAL-SPACE IMPLEMENTATION OF NONLOCAL PSEUDOPOTENTIALS FOR 1ST-PRINCIPLES TOTAL-ENERGY CALCULATIONS
272 KRESSE G,PHYS REV B,vol 0048,page 3115,1993,cites= 16,AB-INITIO MOLECULAR-DYNAMICS FOR OPEN-
SHELL TRANSITION-METALS

273 KRESSE G,PHYS REV B,vol 0049,page 4251,1994,cites= 30,AB-INITIO MOLECULAR-DYNAMICS SIMULATION OF THE LIQUID-METAL AMORPHOUS-SEMICONDUCTOR TRANSITION IN GERMANIUM

SURFACE SCIENCE

DIAMOND SURFACES

274 IARLORI S,PHYS REV L,vol 0069,page 2947,1992,cites= 21,RECONSTRUCTION OF THE DIAMOND %111< SURFACE

275 HIMPSEL FJ,PHYS REV B,vol 0024,page 7270,1981,cites= 19,SURFACE-STATES ON RECONSTRUCTED DIAMOND %111<

276 HAMZA AV,SURF SCI,vol 0237,page 0035,1990,cites= 22,HYDROGEN CHEMISORPTION AND THE STRUCTURE OF THE DIAMOND C%100<-%2X1< SURFACE

277 THOMAS RE,J VAC SCI A,vol 0010,page 2451,1992,cites= 16,THERMAL-DESORPTION FROM HYDROGENATED AND OXYGENATED DIAMOND %100< SURFACES

278 BRENNER DW,PHYS REV B,vol 0042,page 9458,1990,cites= 40,EMPIRICAL POTENTIAL FOR HYDROCARBONS FOR USE IN SIMULATING THE CHEMICAL VAPOR-DEPOSITION OF DIAMOND FILMS

279 GARRISON BJ,SCIENCE,vol 0255,page 0835,1992,cites= 17,MOLECULAR-DYNAMICS SIMULATIONS OF DIMEROPENING ON A DIAMOND %001<%2X1< SURFACE

280 SKOKOV S,J PHYS CHEM,vol 0098,page 7073,1994,cites= 18,ELEMENTARY REACTION-MECHANISM FOR GROWTH OF DIAMOND%100< SURFACES FROM METHYL RADICALS

281 HARRIS SJ,APPL PHYS L,vol 0056,page 2298,1990,cites= 31,MECHANISM FOR DIAMOND GROWTH FROM METHYL RADICALS

EPITAXIAL SURFACE GROWTH

282 TSUNO T,JPN J A P 1,vol 0030,page 1063,1991,cites= 23,EPITAXIALLY GROWN DIAMOND %001< 2X1/1X2 SURFACE INVESTIGATED BY SCANNING TUNNELING MICROSCOPY IN AIR

283 BADZIAN A,DIAM RELAT,vol 0002,page 0147,1993,cites= 11,DIAMOND HOMOEPITAXY BY CHEMICAL-VAPOR-
DEPOSITION

284 HOEVEN AJ,PHYS REV L,vol 0063,page 1830,1989,cites= 13,SCANNING-TUNNELING-MICROSCOPY STUDY OF SINGLE-DOMAIN SI%001< SURFACES GROWN BY MOLECULAR-BEAM EPITAXY

285 HAMERS RJ,ULTRAMICROS,vol 0031,page 0010,1989,cites= 15,NUCLEATION AND GROWTH OF EPITAXIAL SILICON ON SI%001< AND SI%111< SURFACES BY SCANNING TUNNELING MICROSCOPY

286 TSONG TT,REP PR PHYS,vol 0051,page 0759,1988,cites= 11,EXPERIMENTAL STUDIES OF THE BEHAVIOR OF SINGLE ADSORBED ATOMS ON SOLID-SURFACES

287 KELLOGG GL,SURF SCI R,vol 0021,page 0001,1994,cites= 28,FIELD-ION MICROSCOPE STUDIES OF SINGLE-
ATOM SURFACE-DIFFUSION AND CLUSTER NUCLEATION ON METAL-SURFACES

288 MICHELY T,PHYS REV L,vol 0070,page 3943,1993,cites= 27,INVERSION OF GROWTH SPEED ANISOTROPY IN 2-
DIMENSIONS

289 BRUNE H,NATURE,vol 0369,page 0469,1994,cites= 16,MECHANISM OF THE TRANSITION FROM FRACTAL TO DENDRITIC GROWTH OF SURFACE AGGREGATES

290 HWANG RQ,PHYS REV L,vol 0067,page 3279,1991,cites= 22,FRACTAL GROWTH

OF 2-DIMENSIONAL ISLANDS - AU ON RU%0001<
291 BOTT M,SURF SCI,vol 0272,page 0161,1992,cites= 24,THE HOMOEPITAXIAL GROWTH OF PT ON PT%111< STUDIED WITH STM
292 ESCH S,PHYS REV L,vol 0072,page 0518,1994,cites= 16,ORIGIN OF OXYGEN-INDUCED LAYER-BY-LAYER GROWTH IN HOMOEPITAXY ON PT%111<
293 VANDERVEGT HA,PHYS REV L,vol 0068,page 3335,1992,cites= 34,SURFACTANT-INDUCED LAYER-BY-LAYER GROWTH OF AG ON AG%111<
294 COPEL M,PHYS REV L,vol 0063,page 0632,1989,cites= 69,SURFACTANTS IN EPITAXIAL-GROWTH
295 COPEL M,PHYS REV B,vol 0042,page 1682,1990,cites= 33,INFLUENCE OF SURFACTANTS IN GE AND SI EPITAXY ON SI%001<

QUANTUM DOTS

296 EAGLESHAM DJ,PHYS REV L,vol 0064,page 1943,1990,cites= 91,DISLOCATION-FREE STRANSKI-KRASTANOW GROWTH OF GE ON SI%100<
297 LEONARD D,APPL PHYS L,vol 0063,page 3203,1993,cites= 113,DIRECT FORMATION OF QUANTUM-SIZED DOTS FROM UNIFORM COHERENT ISLANDS OF INGAAS ON GAAS-SURFACES
298 BENISTY H,PHYS REV B,vol 0044,page 0945,1991,cites= 48,INTRINSIC MECHANISM FOR THE POOR LUMINESCENCE PROPERTIES OF QUANTUM-BOX SYSTEMS
299 BOCKELMANN U,PHYS REV B,vol 0042,page 8947,1990,cites= 38,PHONON-SCATTERING AND ENERGY RELAXATION IN 2-DIMENSIONAL, ONE-DIMENSIONAL, AND ZERO-DIMENSIONAL ELECTRON GASES
300 BRUNNER K,PHYS REV L,vol 0069,page 3216,1992,cites= 32,PHOTOLUMINESCENCE FROM A SINGLE GAAS/ALGAAS QUANTUM DOT
301 ZRENNER A,PHYS REV L,vol 0072,page 3382,1994,cites= 23,QUANTUM DOTS FORMED BY INTERFACE FLUCTUATIONS IN ALAS/GAAS COUPLED-QUANTUM-WELL STRUCTURES
302 BRUNNER K,APPL PHYS L,vol 0064,page 3320,1994,cites= 20,SHARP-LINE PHOTOLUMINESCENCE OF EXCITONS LOCALIZED AT GAAS/ALGAAS QUANTUM-WELL INHOMOGENEITIES

OPTICS

QUANTUM WELLS

303 BASTARD G,PHYS REV B,vol 0029,page 7042,1984,cites= 12,LOW-TEMPERATURE EXCITON TRAPPING ON INTERFACE DEFECTS IN SEMICONDUCTOR QUANTUM WELLS
304 YANG F,PHYS REV L,vol 0070,page 0323,1993,cites= 10,ORIGIN OF THE STOKES SHIFT - A GEOMETRICAL MODEL OF EXCITON SPECTRA IN 2D SEMICONDUCTORS
305 KOPF RF,APPL PHYS L,vol 0058,page 0631,1991,cites= 10,PHOTOLUMINESCENCE OF GAAS QUANTUM-WELLS GROWN BY MOLECULAR-BEAM EPITAXY WITH GROWTH INTERRUPTIONS
306 GAMMON D,PHYS REV L,vol 0067,page 1547,1991,cites= 12,EXCITONS, PHONONS, AND INTERFACES IN GAAS/ALAS QUANTUM-WELL STRUCTURES
307 KOHL M,PHYS REV B,vol 0039,page 7736,1989,cites= 8,OPTICAL INVESTIGATION OF THE EXCITON TRANSFER BETWEEN GROWTH ISLANDS OF DIFFERENT WELL WIDTHS IN GAAS/ALXGA1-X AS QUANTUM WELLS
308 DEVEAUD B,APPL PHYS L,vol 0051,page 0828,1987,cites= 8,DYNAMICS OF EXCITON TRANSFER BETWEENMONOLAYER-FLAT ISLANDS IN SINGLE QUANTUM WELLS
309 HEGARTY J,PHYS REV B,vol 0030,page 7346,1984,cites= 15,LOCALIZED AND

DELOCALIZED TWO-DIMENSIONAL EXCITONS IN GAAS-ALGAAS MULTIPLE-QUANTUM-WELL STRUCTURES
310 HILLMER H,PHYS REV B,vol 0039,page 0901,1989,cites= 13,OPTICAL INVESTIGATIONS ON THE MOBILITY OF TWO-DIMENSIONAL EXCITONS IN GAAS/GA1-XALXAS QUANTUM WELLS
311 OBERHAUSER D,PHYS REV B,vol 0047,page 6827,1993,cites= 7,EXCITON SCATTERING IN QUANTUM-WELLS AT LOW-TEMPERATURES
312 LUGLI P,APPL PHYS L,vol 0050,page 1251,1987,cites= 6,MONTE-CARLO ALGORITHM FOR HOT PHONONS IN POLAR SEMICONDUCTORS
313 AMAND T,PHYS REV B,vol 0050,page 1624,1994,cites= 7,EXCITON FORMATION AND HOLE-SPIN RELAXATION IN INTRINSIC QUANTUM-WELLS
314 TAKAGAHARA T,PHYS REV B,vol 0031,page 6552,1985,cites= 13,LOCALIZATION AND ENERGY-TRANSFER OF QUASI-2-DIMENSIONAL EXCITONS IN GAAS-ALAS QUANTUM-WELL HETEROSTRUCTURES
315 DAMEN TC,PHYS REV B,vol 0042,page 7434,1990,cites= 32,DYNAMICS OF EXCITON FORMATION AND RELAXATION IN GAAS QUANTUM-WELLS
316 MARTINEZPASTOR J,PHYS REV B,vol 0047,page 0456,1993,cites= 13,TEMPERATURE-DEPENDENCE OF EXCITON LIFETIMES IN GAAS/ALXGA1-XAS SINGLE QUANTUM-WELLS
317 ANDREANI LC,SOL ST COMM,vol 0077,page 0641,1991,cites= 39,RADIATIVE LIFETIME OF FREE-EXCITONS IN QUANTUM-WELLS
318 CITRIN DS,IEEE J Q EL,vol 0030,page 0997,1994,cites= 15,CONTROLLED EXCITON SPONTANEOUS EMISSION IN OPTICAL-MICROCAVITY-EMBEDDED QUANTUM-WELLS
319 HOUDRE R,PHYS REV L,vol 0073,page 2043,1994,cites= 30,MEASUREMENT OF CAVITY-POLARITON DISPERSION CURVE FROM ANGLE-RESOLVED PHOTOLUMINESCENCE EXPERIMENTS
320 WEISBUCH C,PHYS REV L,vol 0069,page 3314,1992,cites= 68,OBSERVATION OF THE COUPLED EXCITON-
PHOTON MODE SPLITTING IN A SEMICONDUCTOR QUANTUM MICROCAVITY
321 ZHU YF,PHYS REV L,vol 0064,page 2499,1990,cites= 28,VACUUM RABI SPLITTING AS A FEATURE OF LINEAR-
DISPERSION THEORY - ANALYSIS AND EXPERIMENTAL-OBSERVATIONS
322 THOMPSON RJ,PHYS REV L,vol 0068,page 1132,1992,cites= 26,OBSERVATION OF NORMAL-MODE SPLITTING FOR AN ATOM IN AN OPTICAL CAVITY

QUANTUM FIELD THEORY

323 REMPE G,PHYS REV L,vol 0058,page 0353,1987,cites= 52,OBSERVATION OF QUANTUM COLLAPSE ANDREVIVAL IN A ONE-ATOM MASER
324 BRUNE M,PHYS REV L,vol 0065,page 0976,1990,cites= 27,QUANTUM NONDEMOLITION MEASUREMENT OF SMALL PHOTON NUMBERS BY RYDBERG-ATOM PHASE-SENSITIVE DETECTION
325 BRUNE M,PHYS REV A,vol 0045,page 5193,1992,cites= 41,MANIPULATION OF PHOTONS IN A CAVITY BY DISPERSIVE ATOM-FIELD COUPLING - QUANTUM-NONDEMOLITION MEASUREMENTS AND GENERATION OF SCHRODINGER CAT STATES
326 YURKE B,PHYS REV L,vol 0057,page 0013,1986,cites= 46,GENERATING QUANTUM-MECHANICAL SUPERPOSITIONS OF MACROSCOPICALLY DISTINGUISHABLE STATES VIA AMPLITUDE DISPERSION
327 DAVIDOVICH L,PHYS REV L,vol 0071,page 2360,1993,cites= 17,QUANTUM SWITCHES AND NONLOCAL MICROWAVE FIELDS
328 CALDEIRA AO,PHYS REV A,vol 0031,page 1059,1985,cites= 15,INFLUENCE OF DAMPING ON QUANTUM INTERFERENCE - AN EXACTLY SOLUBLE MODEL

329 ZUREK WH,PHYS REV D,vol 0026,page 1862,1982,cites= 44,ENVIRONMENT-INDUCED SUPER-SELECTION RULES
330 GRIFFITHS RB,J STAT PHYS,vol 0036,page 0219,1984,cites= 31,CONSISTENT HISTORIES AND THE INTERPRETATION OF QUANTUM-MECHANICS

PHYSICS

ASTROPHYSICS

331 HARTLE JB,PHYS REV D,vol 0028,page 2960,1983,cites= 52,WAVEFUNCTION OF THE UNIVERSE

page 48

Discovering Semantic Patterns in Bibliographically Coupled Documents

JIAN QIN

ABSTRACT

ISSUES IN DISCOVERING KNOWLEDGE IN BIBLIOGRAPHIC databases are addressed. An example of semantic pattern analysis is used to demonstrate the methodological aspects of knowledge discovery in bibliographic databases. The semantic pattern analysis is based on the keywords selected from the documents grouped by bibliographical coupling. The frequency distribution patterns suggest the existence of a common intellectual base with a wide range of specialties and marginal areas in the antibiotic resistance literature. The resulting values for keyword density per rank show a difference of ten times between the specialty and marginal keyword densities. The possibilities and further studies of incorporating knowledge discovery results into information retrieval are discussed.

INTRODUCTION

Knowledge discovery in databases (KDD) is considered a process of nontrivial extraction of implicit, previously unknown, and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases (Chen, Han, & Yu, 1996, p. 866). Most research on KDD has focused on applications in business operations and well-structured data. Knowledge discovery in textual databases has been underemphasized (Trybula, 1997). Among the limited publications on KD in textual databases, the full-text document data are the primary source of analysis. Lent, Agrawal, and Srikant (1997) developed a patent mining system at IBM for identifying trends in large textual databases over a period of time. They

Jian Qin, School of Information Studies, Syracuse University, 4-206 Center for Science and Technology, Syracuse, NY 13244
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 109-132

used sequential pattern mining to identify recurring phrases and generate histories of phrases, after which they then extracted phrases that satisfied a specific trend. Discovering associations among the keywords in texts is another area of research in KD in textual databases. Using background knowledge about the relationships of keywords, Feldman and Hirsh (1996) studied associations among the keywords or concepts representing the documents. The knowledge base they built supplies unary or binary relations among the keywords representing the documents. Feldman, Dagan, and Hirsh (1998) developed a system for Knowledge Discovery in Text (KDT) that extracts keywords to represent document contents and allows users to browse a list of keywords that co-occur with another keyword(s) for knowledge discovery purposes.

Mining in full-text documents attempts to extract useful associations and patterns for representing the document content, including clustering, categorization, summarization, and feature extraction. While many studies using data from bibliographic databases were not conducted in terms of KDD or data mining, they nevertheless bear the marks of KDD's techniques and analysis. Such examples can be found in citation and cocitation analysis (Kassler, 1965; Small, 1973; Small & Sweeney, 1985; Braam, Moed, & van Raan, 1991), keyword classifications (Sparck Jones & Jackson, 1970), investigation of indexing similarities between keywords and controlled vocabularies (Shaw, 1990; Qin, in press), and author mapping (Logan & Shaw, 1987). Discovering knowledge through mining textual data in bibliographic databases presents more problems than mining numerical data. One problem is that most fields in a bibliographic database have long character strings—e.g., author name, title, affiliation, journal title, and indexing terms (from both keywords and controlled vocabularies). Such long strings are usually difficult for statistical packages or data mining software to perform computational tasks. Unlike the full-text document source, bibliographic data are semi-structured. Although it may be an advantage over completely unstructured full-text documents, it also creates a challenge for mining tools that the data in the structured fields should not be mixed up when extracting data sets and performing analysis. Linguistic problems (such as singulars and plurals, stems and suffixes) and inconsistencies in abbreviating journal titles and institution names can also be challenging issues in mining bibliographic data. To obtain valid and reliable data for discovering trends and patterns in subject fields and research, data preprocessing and cleansing can become very time-consuming and both labor and intellectually intensive. However, the most challenging issue remains whether there is a chance for information retrieval systems to "be extended to become knowledge discovery systems," or whether "the kinds of record existing in bibliographical and textual databases offer any possibility of analysis in ways similar to those in more structured factual databases" (Vickery, 1997, pp. 119-20).

This study selected a set of bibliographic records as the data source for discovering semantic patterns among the keywords in these records. The purpose of this keyword analysis was to discover if any semantic patterns existed in the keywords extracted from bibliographically coupled documents regarding antibiotic resistance in pneumonia.

Also, if such patterns did exist, how the discovered knowledge about a subject field can be used to improve the effectiveness of knowledge representation and information retrieval. A preliminary test of antibiotic resistance in pneumonia literature found that documents citing the same publication not only co-cited other publications but also contained semantically similar or same keywords in the titles of cited publications. The frequency distributions of these keywords characterized three distinctive strata: a very small number of keywords falling into the highest frequency region, a relatively larger group with moderate occurrences, and a majority of them appearing only once or twice. If the terms occurring most frequently represent the intellectual base in this subject area (Small, 1973; Small & Sweeney, 1985) and the ones with medium occurrences represent the specialties, then the terms occurring least frequently represent the marginal terms. These marginal terms may be the links between the mainstream of the antibiotic resistance research to the less overt but promising research. The citation-semantic analysis is aimed at discovering semantic patterns of the antibiotic resistance literature so that the analysis process and semantic patterns can be programmed into tools that can assist information searchers in building search queries and customizing their post-search analysis. Specifically, this project studied whether the distribution follows the three strata described earlier, how such distribution can be measured, and to what extent the keywords in these strata reflect the research front in antibiotic resistance. The methods used to preprocess and analyze the data are discussed in detail in the following sections.

Research Design

The first and most important step in KDD is to clarify what kinds of knowledge are to be discovered, because this decides what types of data or database one needs to work on and what techniques to use for discovering the knowledge anticipated. In general, mining data in any type of database includes association rule generalization, multilevel data characterization, data classification, data clustering, pattern-based similarity search, and mining path traversal patterns (Chen, Han, & Yu, 1996). This project was to identify semantic patterns in antibiotic resistance literature, which would be based on the frequency analysis of keyword occurrences. To achieve this goal, one can obtain a set of working data either by selecting keywords directly from individual records or by obtaining a more coherent pool(s) of source documents by applying a citation restriction such as bibliographical coupling. When the bibliographical coupling method is

used to select source documents, at least one similar publication is cited in all the source documents of a bibliographical coupling pool. By this criterion, the documents can be considered coherent in content. Because of this, the keyword data were collected from pools of source documents through bibliographical coupling.

Data Collection

The *Science Citation Index* (*SCI*) database was used to collect data. The following search query was formulated to achieve relative precision and recall:

SELECT (ANTIBIOTIC? (W) RESISTAN?) AND PNEUMONI?

The query was executed in May 1996 and resulted in a total of 360 postings. After ranking by CR (Cited Reference) field, the number of records was reduced to 340 due to the fact that some records did not include references. In Figure 1, these articles are represented by a_1, a_2, a_3, . . . , a_n. A total of 8,753 publications (c_1, c_2, c_3, . . ., c_k in Figure 1) were cited in 340 papers. The highest frequency that a paper was cited was seventy-two times, which means the largest pool of source documents identified via bibliographical coupling contained seventy-two articles (see Table 1). The pools with the same number of source documents were treated as the same rank. All thirty-three ranks in this data set were grouped into three categories: 1 through 10 were large pools, those from 11 to 20 the medium, and the rest the small. The first five pools of source documents were selected from each category for extracting keyword data because of the time constraints for the project. Separate keyword files (i.e., w_1, w_2, w_3, . . ., w_j, in Figure 1) were downloaded for each pool of documents.

Table 1.
Top 10 Most Frequently Cited Documents in Antibiotic Resistance in Pneumonia Literature

Rank	*Frequency of Being Cited*	*Author Name and Source*
1	72	KLUGMAN KP, 1990, V3, P171, CLIN MICROBIOL REV
2	45	MARTON A, 1991, V163, P542, J INFECT DIS
3	41	JACOBS MR, 1978, V299, P735, NEW ENGL J MED
4	38	FENOLL A, 1991, V13, P56, REV INFECT DIS
5	34	HANSMAN D, 1967, V2, P264, LANCET
6	33	APPELBAUM PC, 1992, V15, P77, CLIN INFECT DIS
7	32	SPIKA JS, 1991, V163, P1273, J INFECT DIS
8	28	PALLARES R, 1987, V317, P18, NEW ENGL J MED
9	26	APPELBAUM PC, 1987, V6, P367, EUR J CLIN MICRO
9	26	WARD J, 1981, V3, P254, REV INFECT DIS
10	25	JORGENSEN JH, 1990, V34, P2075, ANTIMICROB AGE
10	25	MUNOZ R, 1991, V164, P302, J INFECT DIS
10	25	PHILIPPON A, 1989, V33, P1131, ANTIMICROB AGEN

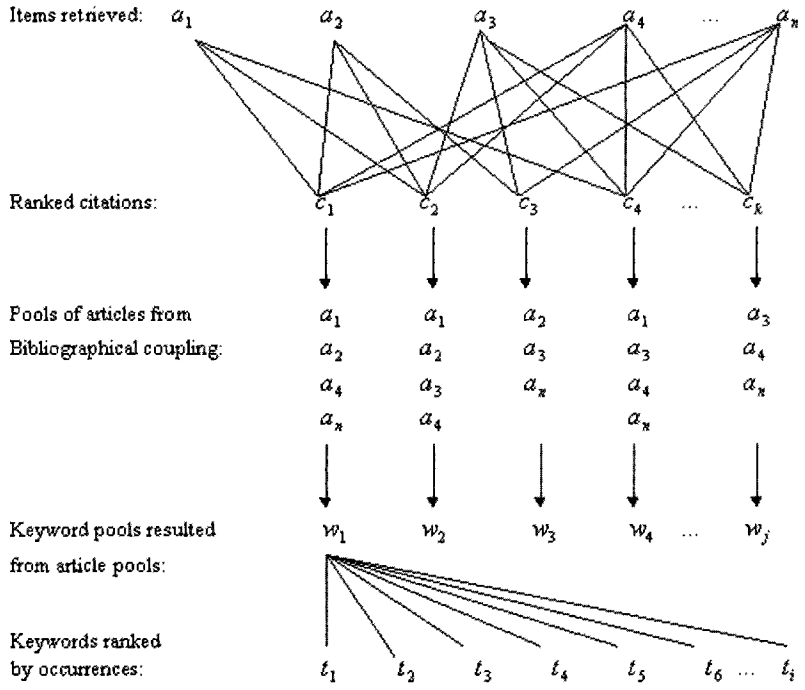


Figure 1. Flowchart of Keyword Extraction.

Data Preprocessing

The first step in preprocessing is cleaning the downloaded keyword files and converting them into tables. This can be done easily with a word processor's FIND and REPLACE functions. Macros or programs can also be written to read the text files into database tables. Data preprocessed through either way would need to be checked for errors, missing values, and the irregular labels missed by the REPLACE command. The next step is then to assign to keywords the text codes that can be computed by analytic tools (see Appendix). As mentioned earlier, textual data mining faces difficulty in handling long character strings and normalizing terms linguistically. Long strings would not be suitable for calculating frequencies or performing other statistical analysis. The text codes designed for the keywords in this subject field are mnemonic and, in most cases, comprehensible without the help of the original forms. A dictionary or a knowledge base for linking text codes to their keywords can be built for automatic coding. In coding keywords for this data set, a general rule was made to maintain as much of the original form and semantics of the keywords as possible. Other coding rules were set as follows:

- The same codes were assigned to both singular and plural forms of the same keywords, e.g., invit = invitro activity/invitro activities, child = child/children.
- The same codes were assigned to those having the same stem but different suffixes, e.g., pn-r = penicillin-resistance/penicillin-resistant, ther = therapeutic and therapy.
- Key phrases were coded by the noun with its modifying adjective or noun as a modifier, e.g., pnu-k = klebsiella pneumoniae, pnu-r = resistant pnuemoniae, pnu-s = streptococcus pnuemoniae.
- A few keywords that were semantically the same but morphologically different were given the same code for the purpose of joining those with the same meanings. Only two keywords fell into this category in this data set: child was used for coding child, children, infants, and pediatric patients; and 3rdw for third-world and developing-countries.

The text coding process was done semi-manually since building the initial code dictionary often needs human intelligence to analyze and translate a keyword or phrase into an appropriate code. The coding consistency (i.e., the same keyword is given the same code or vice versa throughout the data set) was double checked by sorting the data in the order of keyword and text code and then the order of text code and keyword.

Data Analysis

Data analysis in KDD processes is associated with data generalization and summarization which "presents the general characteristics or a summarized high-level view over a set of user-specified data in a database" (Chen, Han, & Yu, 1996, p. 866). The semantic patterns of keywords can be generalized from different perspectives—the simple frequency of occurrences and co-occurrences, or the number of unique keywords per rank (frequency), each of which uses a different measure to analyze the data. The simple frequency of occurrences counts how many times a keyword appears in a bibliographic coupling pool. It draws a high-level view of the semantic patterns from keyword frequency distribution. How often a keyword occurred is often decided by the size of the keyword pool. Obviously, the larger a keyword pool is, the more likely it is for a keyword to occur more frequently. When comparing the simple keyword frequency of a large pool of source documents with that of a smaller one, the result can be misleading because of the uneven bases for comparison. A more meaningful and reliable measure would be relative occurrences—i.e., percentage of times that a keyword appears in the total occurrences.

The frequency of co-occurrences is useful for measuring the importance of a keyword in the subject area, but it needs to be used with care. This data set was divided into large, medium, and small groups of source document pools. A complete coordination of all possible co-occurrences

would involve those between groups 1 and 2, 1 and 3, 2 and 3, and among all three. Even though a keyword may appear in two or three groups at the same time, its frequency of occurrences may vary greatly in different groups. There were also large variations in the numbers of total ranks or frequencies of keyword occurrences: thirty-three in the large group, twenty-four in the medium, and eleven in the small. These can lead to an invalid comparison for the same keyword with the same rank number but in different groups. For instance, a keyword ranked at eleven in the large group, which was considered high in its group, would have meant the lowest rank in the small group.

To normalize the frequency of occurrences, a measure of keyword density per rank was used. The keyword density per rank can be interpreted as the ratio of the number of unique keywords to the number of ranks at which the unique keywords occurred. It can be expressed in the following formula:

$$D(t) = \frac{1}{r_i} \sum_{i=1}^{n} t_i \qquad [1]$$

Where $D(t)$ = Average number of keywords $t_1, t_2, \ldots, t_i$ per rank,

r_i = Number of ranks,

$\sum_{i=1}^{n} t_i$ = Total number of unique keywords included from ranks 1 through n.

Figure 2 shows how the keyword density was calculated.

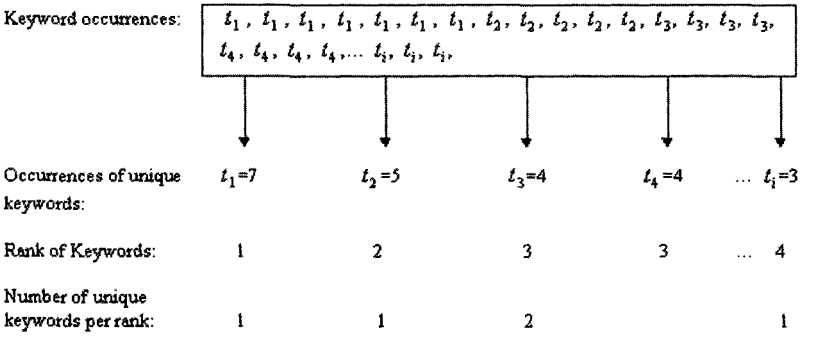


Figure 2. Computation of Keyword Density per Rank.

This measure eliminates the defects of simple frequency and co-occurrences and focuses on how many unique keywords scatter in a region.

This region is denoted by the frequency rank, and its size can be set according to the distribution shape. In Equation [1], the least possible *D(t)* is 1, that is, both the number of unique keywords and the number of ranks are the same. For example, three unique keywords were found to have appeared in three different frequencies (or frequency ranks), then 3/3 = 1. The largest possible *D(t)* can be an infinite in theory, which means that all unique keywords appeared at the same *one* rank. It is clear that the keyword density will increase as a rank contains more unique keywords.

FINDINGS

Frequency Distribution

There were a total of 2,994 keywords in the fifteen pools of source documents. The number of keywords in the large group (source document pools 1-5) consists of 54.5 percent of the total. The medium group had slightly over 40 percent keywords, and the small group only about 10 percent (see Table 2). The decrease in the number of keywords was mainly due to the decrease in the size of document pools; the average number of keywords (7) per record remained approximately the same for each pool. Nonetheless, the frequency distribution of keywords in all three groups was very similar: a majority of the keywords appeared less than five times in each of the groups; as the occurrences increased, the percentage of keywords decreased (see Figure 3).

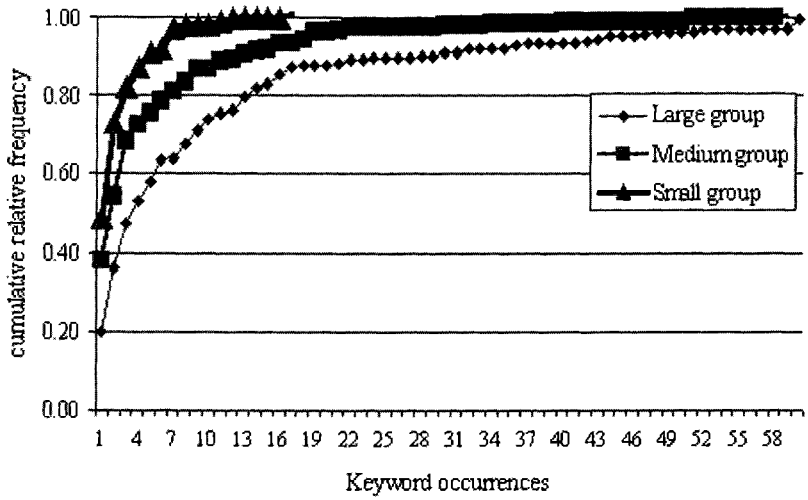


Figure 3. Cumulative Relative Frequency Distribution of Keywords in Three Groups.

Table 2.
NUMBER OF KEYWORDS IN INDEXING RECORDS FOR THE SOURCE DOCUMENTS IDENTIFIED THROUGH BIBLIOGRAPHICAL COUPLING

Group Size by Number of Documents	*Document Pool*	*Number of Documents*	*Number of Keywords*	*Percentage*	*Cumulative Percentage*
Large (Pools of source documents identified through bibliographic coupling)	1	72	512	17.1	17.1
	2	45	316	10.6	27.7
	3	41	291	9.7	37.4
	4	38	273	9.1	46.5
	5	34	241	8.0	54.5
Medium	6	25	200	6.7	61.2
	7	25	208	7.0	68.2
	8	23	207	7.0	75.2
	9	23	171	5.7	80.9
	10	23	202	6.7	87.6
Small	11	11	78	2.6	90.2
	12	11	77	2.6	92.8
	13	10	73	2.4	95.2
	14	10	67	2.2	97.4
	15	10	78	2.6	100.0
Total		401	2,994	100.0	

Although the percentage of keywords declined dramatically as the group size decreased, all three groups shared the same top three keywords—antibiotic resistance, antimicrobial resistance, and streptococcus-pneumoniae. This suggests that a common "intellectual base" existed among all three groups (see Table 3). The percentage of these three keywords dropped in medium and small groups compared to the large group. A close examination of data revealed that the lower occurrences were caused mainly by fluctuations in individual groups (see Figure 4). Figure 4 suggests that such fluctuations became wider as the group size shrunk to the next level.

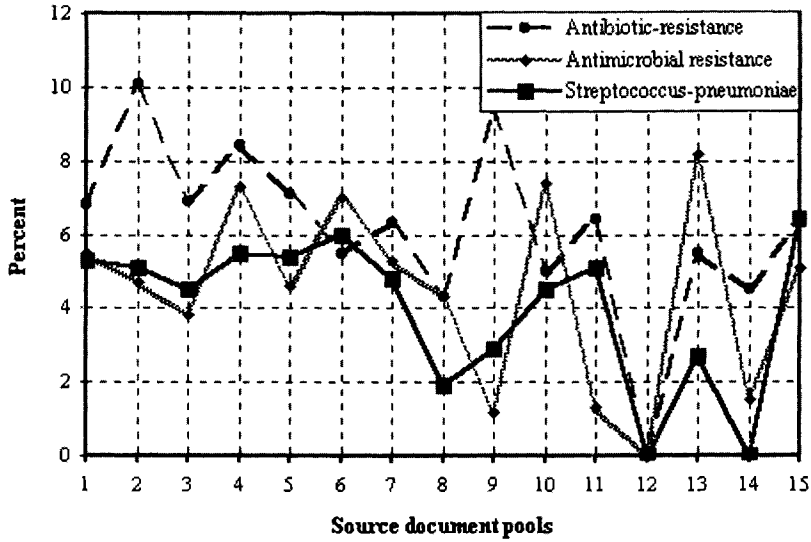


Figure 4. Frequency Distribution of the Intellectual-Base Keywords in 15 Document Pools (1-5 = Large Group, 6-10 = Medium Group, and 11-15 = Small Group).

Co-Occurrence of Keywords

In addition to the base keywords, other keywords co-occurred in either all three groups or two of the three. The largest number of keywords (eighty-five) co-occurred in all three groups. Only seven keywords co-occurred in both the large and small groups besides the eighty-five (see Table 4). The number of unique keywords that occurred in only one group was surprisingly similar: 25, 33, and 33 in the large, medium, and small groups respectively. Among the eighty-five keywords in "all groups" in Table 4, there existed large variations that the same keyword appeared with varied frequencies in different groups. The highest occurrences concentrated in the large group, then declined as the rank of the document pool went down (see Table 5). For example, "children" had sixty-nine occurrences in the large group but decreased to twenty-six and nine respectively in the medium and small groups. While the numbers of unique keywords in the three groups did not differ significantly ($p < 0.05$), the relative occurrences (4.2, 2.6, 2.5 respectively) show that more records in the large group had this keyword. This similar phenomenon happened throughout most other co-occurring keywords in either all three groups or any of the two groups together. Very few keywords that occurred in the small group outnumbered the occurrences in the medium or large group—i.e., although keywords co-occurred in different groups, they did not appear at the same frequency. Keywords co-occurring in only two groups were mostly those with lower frequencies. Figure 4 depicts the number of

Table 3.
RELATIVE FREQUENCIES OF KEYWORDS IN THE FIRST 25TH PERCENTILES IN THREE GROUPS

Keywords	*Large Group*		*Medium Group*		*Small Group*	
	Rank	*Rel. freq.*	*Rank*	*Rel. freq.*	*Rank*	*Rel. freq.*
Antibiotic-resistance	1	7.8	1	6.0	1	4.6
Antimicrobial resistance	2	5.2	2	5.2	2	3.2
Streptococcus-pneumoniae	3	5.1	3	4.0	3	2.9
Children/Infants/Pediatric patients	4	4.2	6	2.6	4	2.2
Susceptibility	5	3.7			6	1.9
Infection/infections					6	1.9
Day-care center/centers			7	2.2		
United States			5	3.0	6	1.9
Haemophilus-influenza			4	3.4		
3rd-generation cephalosporins					5	2.1
Escherichia-co					6	1.9
Mechanically ventilated patients					6	1.9
Penicillin-binding protein			7	2.2		

Table 4.
NUMBER OF UNIQUE KEYWORDS THAT OCCURRED OR CO-OCCURRED IN DIFFERENT GROUPS

	Large Group	*Medium Group*	*Small Group*	*All Groups*
Large Group	25	46	7	
Medium Group	46	33	28	
Small Group	7	28	33	
All Groups	85	85	85	85
Total	163	192	153	

TABLE 5.
PORTION OF THE FREQUENCY AND PERCENTAGE OF THE KEYWORDS THAT CO-OCCURRED

Keywords Occurring in All Three Groups	*Large Group* Freq.	%	*Medium Group* Freq.	%	*Small Group* Freq.	%
Children/Infants/ Pediatric patients	69	4.2	26	2.6	9	2.5
Susceptibility	60	3.7	3	0.3	7	1.9
Pneumococci	47	2.9	21	2.1	4	1.1
Infection/infections	44	2.7	13	1.3	7	1.9
Day-care	42	2.6	22	2.2	5	1.3
United States	37	2.3	6	0.6	7	1.9
Haemophilus-influenza	32	2.0	34	3.4	3	0.8
Therapy/therapeutic	30	1.8	1	0.1	7	1.9
Penicillin resistance	28	1.7	19	1.9	4	1.1
Penicillin-binding protein	24	1.5	22	2.2	5	1.3
Penicillin	22	1.3	6	0.6	1	0.3
Strains	21	1.3	2	0.2	1	0.3
Disease	18	1.1	6	0.6	3	0.8
Epidemiology	17	1.0	11	1.1	2	0.5
Failure	17	1.0	9	0.9	4	1.1
Otitis-media	16	1.0	9	0.9	1	0.3
Vaccine/conjugate vaccine	16	1.0	2	0.2	2	0.5

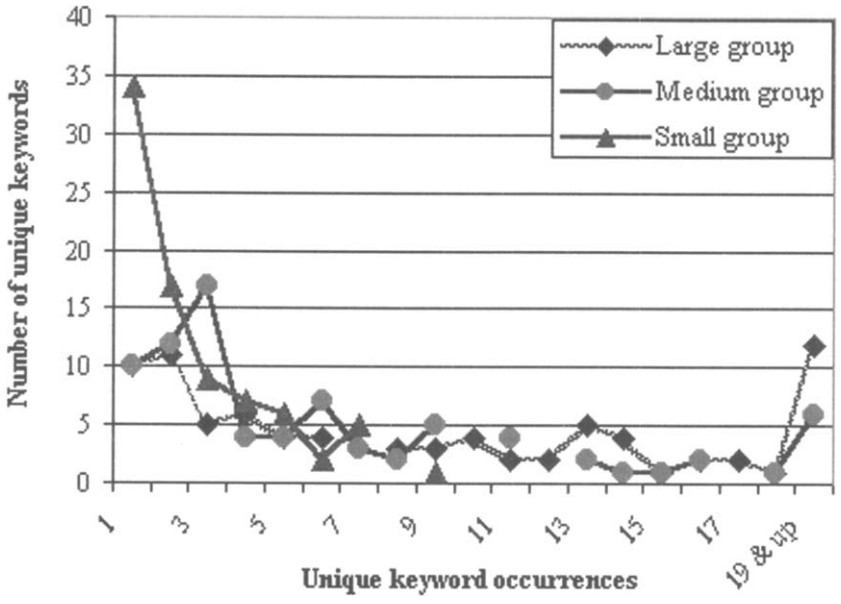


Figure 5. Frequency Distribution of Unique Keywords Occurring in All Three Groups.

unique keywords having occurrences one through more than nineteen. Most unique keywords in the large group occurred more frequently but much less frequently (only one, two, or three times) in the small group.

The Keyword Density

To compute the keyword density per rank, the frequency distribution of keyword occurrences was plotted for each of the three groups after the intellectual base keywords had been excluded. Figure 6 reveals a sharp turn at four, which was then used as a dividing point between the marginal and specialty keywords in the sample. In other words, keywords occurring three or fewer times in the sample were assumed to be marginal in the subject under study, and those with four or more times to be the specialties. Applying Formula [1] in the Data Analysis section, the keyword density was calculated according to the data in Table 6. When calculating the keyword density, ranks that had no keyword occurrences were treated as missing cases and ignored because only the actual number of frequency ranks reflected the keyword density. Thus the number of ranks for specialty keywords in the large group would be 42 minus 3 (intellectual base ranks) minus 3 (marginal ranks) minus 5 (missing cases) equal 31, and so forth for the other two groups. Results in Table 7 show that the density for marginal keywords is approximately ten times greater

than those of specialty keywords in all three groups. Further studies are needed to explore whether this is only a coincidence for this particular data set or a phenomenon existing across disciplines.

Table 6.
FREQUENCY DISTRIBUTION OF KEYWORD OCCURRENCES IN THREE GROUPS EXCLUDING THE THREE INTELLECTUAL BASE KEYWORDS

		Number of Unique Keywords (t_i)					*Number of Unique Keywords (t_i)*		
No.	*Keyword Occurrences*	*Large Group*	*Medium Group*	*Small Group*	*No.*	*Keyword Occurrences*	*Large Group*	*Medium Group*	*Small Group*
1	1	33	72	83	22	24	1		
2	2	26	30	24	23	28	1		
3	3	18	26	15	24	30	1	1	
4	4	10	9	9	25	32	2		
5	5	8	5	7	26	34		1	
6	6	9	8	2	27	36	1		
7	7	1	4	8	28	37	1		
8	8	6	5	1	29	40		1	
9	9	5	6	1	30	42		1	
10	10	5			31	43	1		
11	11	2	4	1	32	44	1		
12	12	2	2	1	33	47	1		
13	13	5	2		34	48	1		
14	14	4	2		35	51	1		
15	15	1	1		36	52	1		
16	16	4	3		37	59	1		
17	17	3		1	38	60	1		
18	18	1	2		39	69	1		
19	19		3		40	84	1		
20	21	1	1		41	85	1		
21	22	1	2		42	129	1		
						Total	163	192	153

Table 7.
KEYWORD DENSITY IN GROUPS

Keyword Density (D)	*Intellectual Base Keywords (i)*	*Specialty Keywords (s)*	*Marginal Keywords (m)*
Large Group (l)	D(li)=3/3=1	D(ls)=83/31=2.68	D(lm)=77/3=25.67
Medium Group (m)	D(mi)=3/3=1	D(ms)=61/19=3.21	D(mm)=128/3=42.67
Small Group (s)	D(si)=3/3=1	D(ss)=28/6=4.67	D(sm)=122/3=40.67

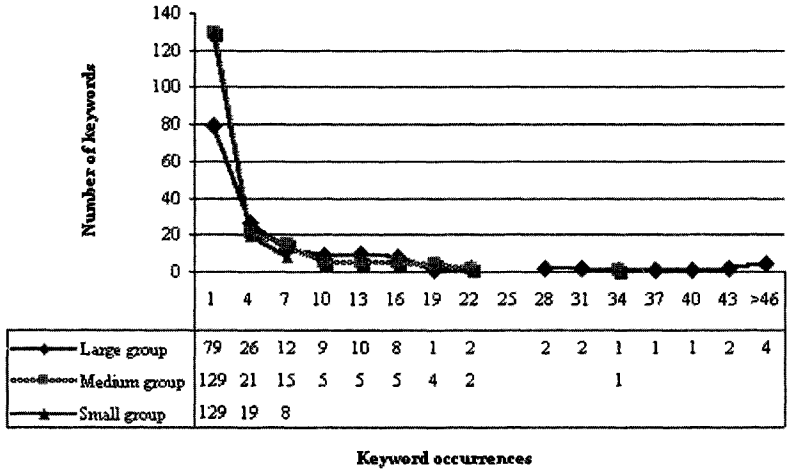


Figure 6. Frequency Distribution of Keyword Occurrences in Three Groups.

A further examination was made for keywords in the specialty and marginal groups. Several patterns emerged in the specialty keywords (see Tables 8, 9, and 10):

- Keywords co-occurring in two or three groups tended to be more generic or disciplinarily generic than non-co-occurring ones. Examples included children, day-care, failure, infections, prevalence, United States, genes.
- There were more microbial names and related infections in the keywords co-occurring than in the ones not co-occurring. Examples included pneumococci, enterococcus/enterococci, Escherichia-coli, Klebsiella-pneumonia, Neisseria-gonorrhoea, haemophilus-influenza, streptococcus-pneumonicoccal meningitis, Branhamella-catarrhalis.
- There was a clear tendency in the keywords both co-occurring and non-co-occurring (the latter happened in the first two groups only) that antibiotic resistance in pneumonia was investigated from perspectives of genetics (binding protein gene, penicillin-binding proteins, multiresistant clone), microbiology (invitro activities), and immunology (pneumococcal polysaccharide). However, this tendency in co-occurring keywords seemed to be more toward pharmaceutical aspects in relation to the microbes and infections they caused (spectrum beta-lactam, chloramphenicol therapy, third-generation cephalosporins), and more pathologically oriented in non-co-occurring keywords (serotype distribution, strains, antimicrobial susceptibility, plasmids).
- The keyword density in double digits were generally more specific than those in single digits, though there did exist a few general ones (see Table 10).

Table 8.
SPECIALTY KEYWORDS THAT CO-OCCURRED IN TWO OR THREE GROUPS

		Large Group		*Medium Group*		*Small Group*	
No.	*Keywords*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
1	Children/Infants/ Pediatric patients	69	4.2	26	2.6	9	2.4
2	Pneumococci	47	2.9	21	2.1	4	1.1
3	Infection/infections	44	2.7	13	1.3	7	1.9
4	Day-care/Day-care centers	42	2.6	22	2.2	5	1.3
5	United States	37	2.3	6	0.6	7	1.9
6	Penicillin resistance	28	1.7	19	1.9	4	1.1
7	Penicillin-binding proteins	24	1.5	22	2.2	5	1.3
8	Failure	17	1.0	9	0.9	4	1.1
9	Gene/genes	14	0.9	15	1.5	4	1.1
10	Pneumococcal polysaccharide	14	0.9	7	0.7	7	1.9
11	Prevalence	13	0.8	11	1.1	5	1.3
12	Enterococcus/ enterococci	6	0.4	16	1.6	4	1.1
13	Escherichia-coli	5	0.3	16	1.6	7	1.9
14	Spectrum beta-lactam	5	0.3	18	1.8	5	1.3
15	Klebsiella-pneumoni	4	0.2	9	0.9	6	1.6
16	Neisseria-gonorrhoea	4	0.2	4	0.4	6	1.6
17	Meningitis	52	3.2	16	1.6		
18	Chloramphenicol therapy	36	2.5	8	0.8		
19	Haemophilus-influenza	32	2.0	34	3.4		
20	Penicillin	22	1.3	6	0.6		
21	Disease	18	1.1	6	0.6		
22	Epidemiology	17	1.0	11	1.1		
23	Binding protein gene	16	1.0	14	1.4		
24	Otitis-media	16	1.0	9	0.9		
25	Streptococcus-pneumonicoccal meningitis	15	0.9	14	1.4		
26	Bacterial-meningitis	14	0.9	11	1.1		
27	Bacteria	13	0.8	6	0.6		
28	Beta-lactam antibiotics	13	0.8	8	0.8		
29	Branhamella-catarrhalis	12	0.7	7	0.7		
30	Multiresistant clone	12	0.7	8	0.8		
31	Antibiotics	11	0.7	9	0.9		
32	Influenzae type-b	11	0.7	11	1.1		
33	Invitro activities	10	0.6	5	0.5		
34	Carriage	9	0.6	4	0.4		
35	Diagnose	9	0.6	4	0.4		

36	Resistance	9	0.6	6	0.6		
37	Erythromycin	8	0.5	4	0.4		
38	Staphylococcus-aureu	8	0.5	11	1.1		
39	Influenzae	4	0.2	5	0.5		
40	Tuberculosis	4	0.2	6	0.6		
41	Susceptibility	60	3.7			7	1.9
42	Therapy/ Therapeutic	30	1.8			7	1.9
43	Emergence	8	0.5			4	1.1
44	Protective efficacy	5	0.3			4	1.1
45	3rd-generation cephalosporins			6	0.6	8	2.2
46	Enterobacter/ Enterobacteriaceae			7	0.7	5	1.3
47	Respiratory-tract infection			4	0.4	4	1.1

Table 9.
SPECIALTY KEYWORDS OCCURRING IN A SINGLE GROUP

Keywords Unique in the Large Group				*Keywords Unique in the Medium*			
No.	*Group*	*Freq.*	*%*	*No.*	*Group*	*Freq.*	*%*
1	Serotype distribution	48	2.9	37	UK	30	3
2	Spain	32	2.0	38	Sulbactam	19	1.9
3	Strains	21	1.3	39	Resistant staphyl-ococcuci	18	1.8
4	New-Guinea/ Papua-New-Guinea	17	1.0	40	Nucleotide-sequences	13	1.3
5	Vaccine/Conjugate vaccine	16	1.0	41	Sri-Lanka	12	1.2
6	Pneumococcal meningitis	14	0.9	42	Cephalosporins	9	0.9
7	Systemic infections	13	0.8	43	Pneumococ-cal serotype	9	0.9
8	Penicillin-resistant pneumoniae	13	0.8	44	Tetracycline	8	0.8
9	Resistant pneumo-coccus pneumoniae	10	0.6	45	Transpeptidase	8	0.8
10	Upper respiratory -tract	10	0.6	46	Mechanism	6	0.6
11	Community-acquired pnuemoniae	10	0.6	47	Catarrhalis beta-lactamase	5	0.5
12	Immune-deficiency syndrome	9	0.6	48	Pneumococcal vaccine	5	0.5
13	Antibody	9	0.6	49	Postsplenectomy sepsis	5	0.5
14	Vancomycin	8	0.5	50	Ampicillin	4	0.4
15	Horizontal transfer	8	0.5	51	Gram-negative bacilli	4	0.4
16	Antimicrobial susceptibility	8	0.5	52	Plasmid/plasmids	4	0.4

continued on page 126

table 9 continued

No.	*Keywords Unique in the Medium Group*	*Freq.*	%	No.	*Keywords Unique in the Small Group*	*Freq.*	%
17	Hungary	7	0.4	53	Mechanically ventilated patients	7	1.9
18	Iceland	6	0.4				
19	Invasive disease	6	0.4	54	Patients/critically ill patients	7	1.9
20	Patterns	6	0.4				
21	Pneumococcal infections	6	0.4	55	Intensive-care unit (AIDS)	5	1.3
22	Acquired immuno defficiency syndrome	6	0.4	56	Nosocomial infection	5	1.3
				57	Transferable resistance	4	1.1
23	Bacterial pneumonia	6	0.4				
24	Cerebrospinal-fluid	6	0.4				
25	Co-trimoxazole	6	0.4				
26	Requiring hospitalization	5	0.3				
27	Sensitivity	5	0.3				
28	Septic arthritis	5	0.3				
29	Human-Immuno deficiency Virus (HIV)	5	0.3				
30	Capsular poly-saccharide	5	0.3				
31	Molecular epidemiology	4	0.2				
32	Invasive pneum-ococcal infections	4	0.2				
33	Anemia	4	0.2				
34	Binding proteins	4	0.2				
35	Capsular types	4	0.2				
36	Ciprofloxacin	4	0.2				

Table 10.
MARGINAL KEYWORDS THAT CO-OCCURRED IN EITHER ALL THREE OR TWO OF THE THREE GROUPS

Keywords	*Large Group*		*Medium Group*		*Small Group*	
	Freq.	%	*Freq.*	%	*Freq.*	%
Pseudomonas-aeruginosa	3	0.2	3	0.3	2	0.5
Management	3	0.2	3	0.3	1	0.3
Etiology	3	0.2	1	0.1	1	0.3
Isolate/clinical isolate	2	0.1	2	0.2	1	0.3
Coagulase-negatives	2	0.1	2	0.2	1	0.3
Aminoglycoside resistance	2	0.1	1	0.1	2	0.6
Legionnaires-disease	2	0.1	1	0.1	1	0.3
Microdilution system	2	0.1	1	0.1	1	0.3
Blood cultures	2	0.1	2	0.2	1	0.3
Trimethoprim-sulfame	1	0.1	2	0.2	1	0.3
Methicillin-resistant	1	0.1	2	0.2	1	0.3

Refractory periodont	1	0.1	1	0.1	1	0.3
Outer-membrane permeability	1	0.1	3	0.3	1	0.3
Outbreak	1	0.1	3	0.3	1	0.3
Nosocomial outbreak	1	0.1	3	0.3	2	0.5
Colonization	1	0.1	3	0.3	3	0.8
2x	2	0.1	2	0.2		
Anti-inflammatory agent	3	0.2	1	0.1		
Antibiotic-therapy	1	0.1	1	0.1		
Aspiration	1	0.1	1	0.1		
Broth	1	0.1	1	0.1		
Cefamandole	3	0.2	1	0.1		
Ceftriaxone	3	0.2	1	0.1		
Clarithromycin	1	0.1	2	0.2		
Clindamycin	1	0.1	1	0.1		
Clones	2	0.1	2	0.2		
Common organization	2	0.1	1	0.1		
D-alanine ligase	2	0.1	3	0.3		
Directions	1	0.1	1	0.1		
Group-a	1	0.1	1	0.1		
High-level resistance	2	0.1	3	0.3		
Invasive pneumococcal infections	1	0.1	1	0.1		
Nasopharyngeal carriage	3	0.2	2	0.2		
Neisseria-meningitis	1	0.1	2	0.2		
Pathogen	1	0.1	1	0.1		
Populations	3	0.2	2	0.2		
Quinolones	1	0.1	1	0.1		
South-Africa	2	0.1	1	0.1		
Spread	3	0.2	1	0.1		
Streptococcus-pneumoniae strains	2	0.1	1	0.1		
Structural-changes	1	0.1	1	0.1		
Ampicillin	3	0.2			1	0.3
Antimicrobial agents	3	0.2			1	0.3
Bacterium legionella	2	0.1			1	0.3
Catarrhalis beta-lactamase	3	0.2			1	0.3
Cephalosporins	2	0.1			3	0.8
Dirithromycin	1	0.1			1	0.3
Mechanism	2	0.1			2	0.5
Norfloxacin	1	0.1			1	0.3
Nucleotide-sequences	2	0.1			2	0.5
Plasmid/plasmids	2	0.1			2	0.5
Pneumococcal vaccine	1	0.1			1	0.3
Spectrum	1	0.1			1	0.3
Affairs-medical-center			2	0.2	1	0.3
Anemia			1	0.1	1	0.3
Antibody			3	0.3	1	0.3
Aztreonam			1	0.1	1	0.3
Broad-spectrum cepha			1	0.1	1	0.3
Calcoaceticus var anitratus			2	0.2	1	0.3
Capsular polysaccharide			3	0.3	1	0.3

continued on page 128

table 10 continued

Keywords	*Large Group* *Freq.*	%	*Medium Group* *Freq.*	%	*Small Group* *Freq.*	%
Ceftazidime resistance			3	0.3	1	0.3
Cerebrospinal-fluid			3	0.3	2	0.5
Ciprofloxacin			3	0.3	1	0.3
Classification			2	0.2	1	0.3
Community-acquired pnuemoniae			2	0.2	2	0.5
Digestive-tract			1	0.1	3	0.8
DNA			3	0.3	3	0.8
Enzymatic resistance			3	0.3	2	0.5
Horizontal transfer			1	0.1	2	0.5
Identification			1	0.1	1	0.3
Imipenem-cilastatin			3	0.3	3	0.8
India			1	0.1	1	0.3
Nursing-home patient			3	0.3	1	0.3
Patterns			1	0.1	1	0.3
Penicillin-resistant pneumoniae			3	0.3	1	0.3
Pneumococcal meningitis			3	0.3	3	0.8
Resistant pneumococcus pneumoniae			3	0.3	1	0.3
Salmonella-typhi			1	0.1	1	0.3
Selective decontamination			2	0.2	3	0.8
Staphylococcus-aureu			3	0.3	2	0.5
Steady-state treatment			2	0.2	1	0.3
Strains			2	0.2	1	0.3
Substitution			1	0.1	1	0.3
Systemic infections			2	0.2	1	0.3
Third-world countries			1	0.1	2	0.5
Transposition			1	0.1	1	0.3
Upper respiratory-tract			2	0.2	3	0.8
Vancomycin			1	0.1	1	0.3

CONCLUSION

Knowledge discovery in bibliographic databases is distinctive compared to KD in full-text document and numerical databases. One challenge is transforming semi-structured textual data into the types and structures suitable for calculations and modeling. In the case of subject keywords, all the idiosyncrasies existing in natural language, including suffixes, different spellings for the same word, and synonyms, need to be normalized before analysis.

Similar work on this type of term normalization has been done in automatic indexing, such as stem stripping (Paice, 1990; Porter, 1980). Harman and Candela (1990) argue that term normalization such as stem stripping is not worth the effort for large full-text databases because this operation has little impact on other methods (e.g., frequency counts) of

indexing. While this may be true in indexing full-text documents, preprocessing of this kind is a necessity in discovering knowledge from bibliographic databases. The reason is obvious: semantic analysis of subject keywords needs to have accurate data to draw reliable and valid conclusions. The unnormalized keyword data set can present false patterns or trends in keywords. Although term normalization is a time-consuming operation, it can be improved by making use of the prior research and database technology mentioned earlier. In this study, as the initial text code base took shape, coding became easier and quicker as more and more text codes were established. It was found that semantic coding, while resembling vocabulary control, is different from vocabulary control in indexing. Semantic coding groups semantically same and/or similar keywords together by using simple codes that can be easily constructed. Using these codes, original terms can be preserved with semantic value-added processing, thus there is no need for using totally different "controlled" terms to substitute the keywords used in the publications. It quantifies the information analysis process and turns the jobs requiring expert knowledge into relatively simple tasks. This method is particularly suitable for specialized and interdisciplinary subject fields.

Presentation of the knowledge discovered is an important part of the KDD process. Visualizing the patterns, trends, and associations in a subject field can be very challenging because of the size of the screen and the number of text values that one data field can contain. This study of semantic patterns in keywords was by no means a large one in scale, but the total number of keywords made it difficult to draw any legible charts for the whole data set. An inclusion of even one group of keywords would clutter the chart badly and cause the keywords on the chart to be unrecognizable. Substituting long keywords with shorter and mnemonic text codes normalized the inconsistencies in keywords as well as leaving more room for visual presentation of the knowledge discovered.

The semantic patterns discovered in this data set suggest that different keyword density regions may be used as a controlling mechanism for better targeted searching. Traditionally, query expansion is one of the main techniques used to improve retrieval performance (Sparck Jones & Jackson, 1970; Salton, Fox, & Voorhees, 1983; Salton & Buckley, 1990; Harman, 1992). Query expansion allows searchers to browse the indexing term list or give relevance feedback to searchers through frequency ranking or term weighting. While performance was reported to have improved to a high percentage in small testing collections, it is still unproven that these techniques would achieve the same performance in large collections in the real world (Korfhage, 1997). Keyword density may provide a new solution to this uncertainty because it is computed on the basis of a collection of keywords extracted from bibliographically

coupled source documents. By implementing keyword density analysis into an algorithm, it is possible that a simple search query to the database(s) would generate a group of keywords stratified by their density regions. Information searchers can then select keywords from different density regions according to their own definition of relevance.

The semantic patterns found in non-co-occurring and co-occurring keywords suggest that it is necessary and possible to design new search tools that will deliver "analyzed" search results to users. In retrieving information from terabyte databases, the most challenging task in information retrieval is probably how to find most relevant information, in a manageable amount, in the easiest way. Conventional information systems have applied various sophisticated methods to accomplish this task but were limited by their design which requires equally sophisticated search techniques to find the information and leaves the information filtration to users themselves. The development of science and the growth of scientific literature has made filtering relevant information more difficult than ever in highly interdisciplinary scientific research areas. The semantic pattern analysis of the keywords from bibliographical coupling shows a possibility that simple semantic processing to natural language (keywords extracted from citations in this case) may be programmed and serve as a tool for providing "analyzed" search results to users.

The results in this study are only preliminary. It is unknown whether the semantic patterns identified in this data set are a coincidence or a common phenomenon across subject fields. Further studies are needed to discover whether the subject category of keywords is related to the density region and whether the stratified keyword distribution and density can contribute to customizing the selection of a targeted group of documents and post-search analysis.

APPENDIX

EXAMPLES OF KEYWORDS AND THEIR SEMANTIC CODES IN THE SAMPLE

Keyword	*Code*	*Keyword*	*Code*
2x	2x	HUMAN-IMMUNO-DEFICIENCY VIRUS	hiv
ANTI-INFLAMM-ATORY AGENT	agnt	HUNGARY	hungary
AID	Saids	IMMUNE-DEFICIENCY SYNDROM	Eids
ANTIMICROBIAL SUSCEPTIBILITY	sus-am	INVASIVE PNEUMO-COCCAL INFECTION	Sinf-pnu
ASPIRATION	aspirat	PNEUMOCOCCAL INFECTIONS	inf-pnu
BARCELONA	barc	INVASIVE DISEASE	invasiv
BINDING PROTEINS	bp	MENINGITIS/ MENINGEAL	meningi
BINDING PROTEIN GENE	bpg	MOLECULAR EPIDEMIOLOGY	epi-mol
BROTH	broth	NEW-GUINEA/ PAPUA-NEW	ng
CAPSULAR TYPES	capt	COMMON ORGAN-IZATION	org
CARRIAGE	carrig	PATHOGEN	pathoge
NASOPHARYNGEAL CARRIGE	carr-n	BACTERIAL PNEUMONIA	pnu-b
CEFAMANDOLE	cefam	QUINOLONES	quinolo
CEFTRIAXONE	ceftri	HIGH-LEVEL RESISTANCE	res-hi
CHLORAMPHEN -ICOL THERAPY	ther-chl	SOUTH-AFRICA	sa
CLARITHROMYCIN	clarith	SENSITIVITY	sensiti
CLINDAMYCIN	clindam	POSTSPLE- SEPSIS NECTOMY	sepsi-p
CLONES	clone	SPREAD	spread
MULTIRESISTANT CLONE	clone-m	STREPTOCOCCUS -PNEUMONIAE STRAINS	stra-s
DIRECTIONS	direct	SOUTH-AFRICAN STRAIN	stra-sa
D-ALANINE LIGASE	ligase	STRUCTURAL-CHANGES	struct
ERYTHROMYCIN	erythr	TETRACYCLINE	tetracy
GROUP-A	grp-a	ANTIBIOTIC-THERAPY	ther-a

References

Braam, R. R.; Moed, H. F.; & van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. Part I: Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233-251.

Braam, R. R.; Moed, H. F.; & van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis. Part II: Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252-266.

Chen, M. Y.; Han, J.; & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8,* 866-883.

Feldman, R., & Hirsh, H. (1996). Mining associations in text in the presence of background knowledge. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *KDD '96* (Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, August 1996) (pp. 343-346). Menlo Park, CA: AAAI Press.

Feldman, R.; Dagan, I.; & Hirsh, H. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems, 10*(3), 281-300.

Harman, D. (1992). User-friendly systems instead of user-friendly front-ends: Four end-user systems employing probabilistic ranking: PRISE, CITE, MUSCAT and News Retrieval Tool. *Journal of the American Society for Information Science, 43*(2), 164-174.

Harman, D. K., & Candela, G. (1990). Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *Journal of the American Society for Information Science, 41*(8), 581-589.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation, 16*(3), 223-233.

Korfhage, R. R. (1997). *Information storage and retrieval.* New York: Wiley.

Lent, B.; Agrawal, R.; & Srikant, R. (1997). Discovering trends in text databases. In D. Heckerman, H. Mannila, & D. Pregibon (Eds.), *KDD '97* (Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, August 14-17, 1997) (pp. 227-230). Menlo Park, CA: AAAI Press.

Logan, E. L., & Shaw, W. M. (1987). An investigation of the coauthor graph. *Journal of the American Society for Information Science, 38*(4), 262-268.

Paice, C. (1990). Another stemmer: Natural language processing. *SIGIR Forum, 24*(3), 56-61.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Qin, J. (In press). Indexing similarities in a keyword database and a controlled vocabulary database: Antibiotic resistance in pneumonia. *Journal of the American Society for Information Science.*

Salton, G.; Fox, E. A.; & Voorhees, E. M. (1985). Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science, 36*(2), 200-210.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288-297.

Shaw, W. M. (1990). Subject indexing and citation indexing: Clustering structure in the cystic fibrosis document. *Information Processing and Management, 26*(6), 693-718.

Small, H. (1973). Co-citation in scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations: I. A comparison of methods. *Scientometrics, 7*(3/6), 391-409.

Sparck Jones, K., & Jackson, D. M. (1970). The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval, 5*(4), 175-201.

Travis, J. (1994). Reviving the antibiotic miracle. *Science, 264*(5157), 360-362.

Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology, 32,* 197-229.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Additional Reference

Small, H.; Sweeney, E.; & Greenlee, E. (1985). Clustering the *Science Citation Index* using co-citations: II. Mapping science. *Scientometrics, 8*(5/6), 321-340.

Knowledge Discovery Through Co-Word Analysis

Qin He

Abstract

In the last half century, as the science literature has increased dramatically, scientists found it increasingly difficult to locate needed data, and it is increasingly difficult for policymakers to understand the complex interrelationship of science in order to achieve effective research planning. Some quantitative techniques have been developed to ameliorate these problems; co-word analysis is one of these techniques. Based on the co-occurrence frequency of pairs of words or phrases, co-word analysis is used to discover linkages among subjects in a research field and thus to trace the development of science. Within the last two decades, this technique, implemented by several research groups, has proved to be a powerful tool for knowledge discovery in databases. This article reviews the development of co-word analysis, summarizes the advantages and disadvantages of this method, and discusses several research issues.

Introduction

Since World War II, the scope and volume of scientific research have increased dramatically. This is well reflected in the growth of the literature. In the 1960s, the amount of scientific literature was estimated to be doubling approximately every ten years (Price, 1963). Three decades later, in the 1990s, along with developments in information technology, especially in the area of data storage, the amount of information in the world is estimated to be doubling every twenty months (Frawley et al., 1991). In such a situation, it is hard for scientists to detect the subject areas and the

Qin He, Graduate School of Library and Information Science, University of Illinois, 501 E. Daniel Street, Champaign, IL 61820
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 133-159

linkages among these areas in their research fields, and policy makers have difficulties in mapping the dynamics of science to do research planning.

The traditional way to map the relationships among concepts, ideas, and problems in science is to seek the views of a relatively small number of experts. Even though such methods are indispensable for some purposes, as Law and Whittaker (1992) said, they also have certain drawbacks:

> First, they are extremely expensive unless the survey of experts is very small. Second, if the survey is small, then its representativeness is open to question. Third, the problem of collating a range of views about the way in which science has developed or is developing is complex. (pp. 417-418)

For these reasons, quantitative methods for mapping the structure of science have been developed; they include co-citation analysis, co-nomination analysis, and co-word analysis. This article reviews the development of the co-word analysis technique.

Co-word analysis is a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., words or noun phrases) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these texts. Indexes based on the co-occurrence frequency of items, such as an inclusion index and a proximity index, are used to measure the strength of relationships between items. Based on these indexes, items are clustered into groups and displayed in network maps. For example, an inclusion map is used to highlight the central themes in a domain, and a proximity map is used to reveal the connections between minor areas hidden behind the central ones. Some other indexes, such as those based on density and centrality, are employed to evaluate the shape of each map, which shows the degree to which each area is centrally structured and the extent to which each area is central to the others. By comparing the network maps for different time periods, the dynamic of science can be detected.

The co-word analysis technique was first developed in collaboration between the Centre de Sociologie de l'Innovation of the École Nationale Superieure des Mines of Paris and the CNRS (Centre National de la Recherche Scientifique) of France during the 1980s, and their system was called "LEXIMAPPE." For about twenty years, this technique has been employed to map the dynamic development of several research fields. One of the early studies was carried out by Serge Bauin (1986) to map the dynamics of aquaculture from 1979 to 1981. Based on the inclusion and proximity indexes, inclusion and proximity maps were created for 1979 and 1981.

With the decomposition of keywords into central poles and mediator words, the inclusion map for 1979 is shown in Figure 1 and that for 1981 is shown in Figure 2.

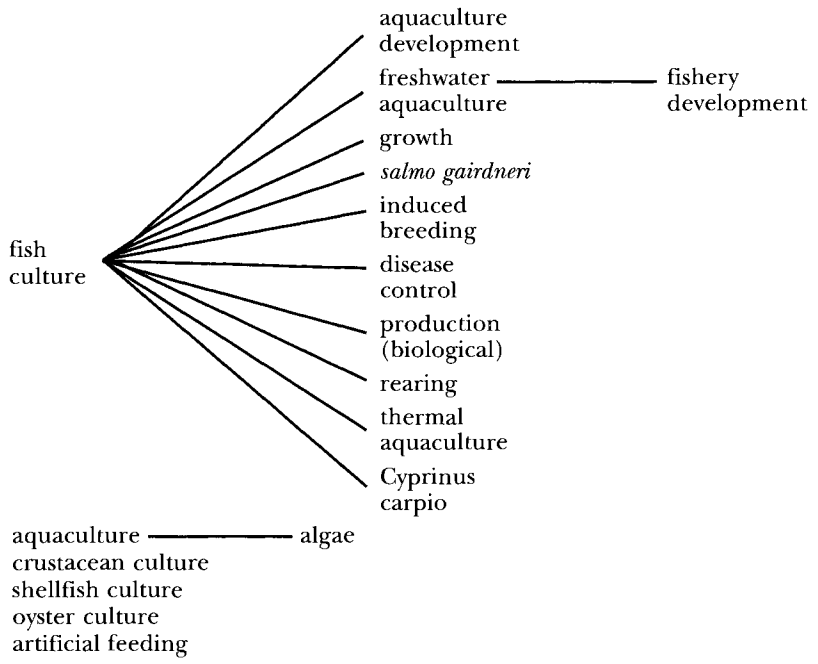


Figure 1. Main Central and Mediator Words, 1979 (Bauin, 1986, p. 127.) Reprinted with kind permission of Macmillan Press, Ltd.

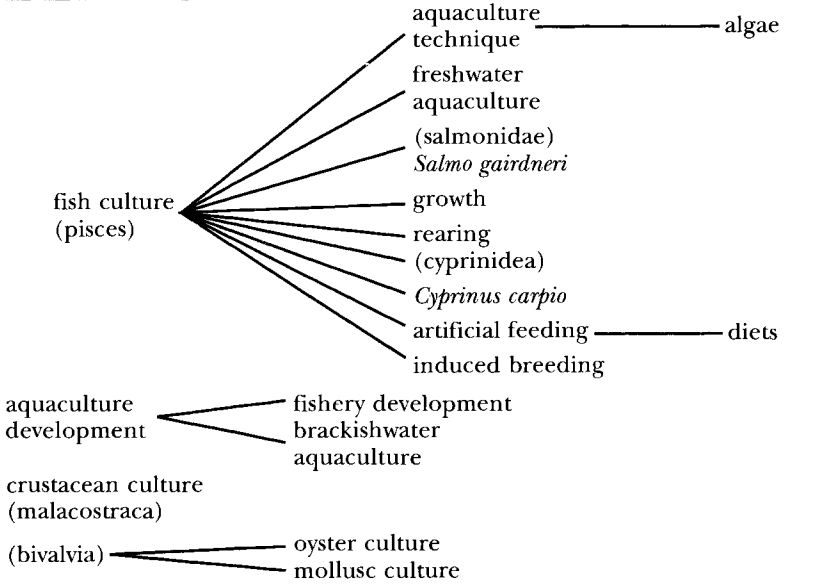


Figure 2. Main Central and Mediator Words, 1981 (Bauin, 1986, p. 128.) Reprinted with kind permission of Macmillan Press, Ltd.

In the map for 1979, "*Salmo gairdneri,*" a fish species which has been bred less extensively in Norway's seas since the 1950s, remained unexpectedly as a high frequency mediator word. However, in the map for 1981, this term was replaced by "salmonidae." One of the more significant changes is that the central pole "aquaculture" in the 1979 map has disappeared. It has been replaced by two new poles—"aquaculture development" and "aquaculture techniques." In addition, the word "artificial feeding" loses its status as a central pole in the map for 1979 and appears under "fish culture" in the map for 1981.

The proximity maps for 1979 and 1981 respectively are shown in Figures 3 and 4.

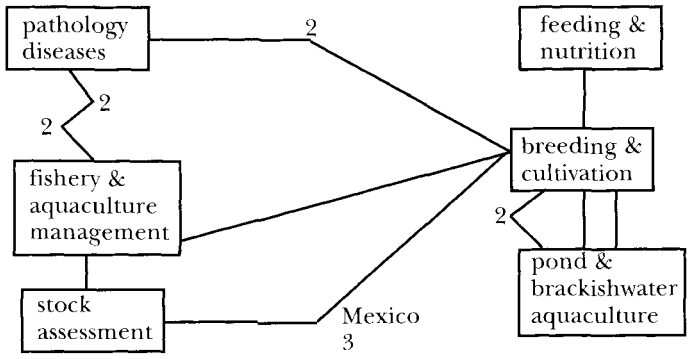


Figure 3. Proximity Map, 1979 (Bauin, 1986, p. 133). Reprinted with kind permission of Macmillan Press, Ltd.

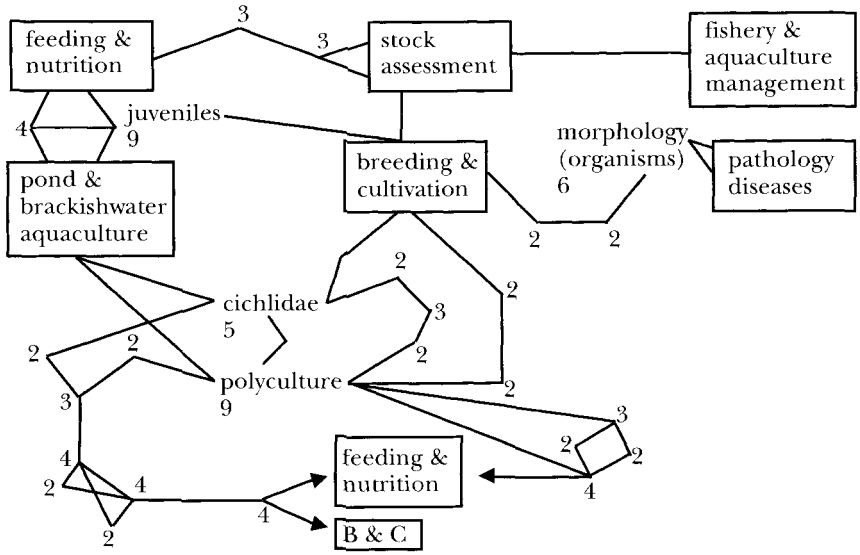


Figure 4. Proximity Map, 1981 (Bauin, 1986, p. 134). Reprinted with kind permission of Macmillan Press, Ltd.

Comparing the two maps, it is noted that, from 1979 to 1981, some clusters, such as "feeding and nutrition," become extended and more structured—i.e., the average number of links per word has increased. Overall, the average number of links per word in the complete maps has increased from 2.33 to 2.95. This might be an indication of the beginning of the integration of the whole field.

This and other examples (e.g., Turner & Callon, 1986; Callon, 1986; Courtial & Law, 1989; Law & Whittaker, 1992; Coulter et al., 1998) reveal that co-word analysis is a promising method for discovering associations among research areas in science and for revealing significant linkages that may otherwise be difficult to detect. It is a powerful tool that makes it possible to trace the structure and evolution of a socio-cognitive network (Bauin, 1986). As such, it offers a significant approach to knowledge discovery.

The Development of Co-Word Analysis

In 1986, Callon, Law, and Rip (1986) edited a book titled *Mapping the Dynamics of Science and Technology*. This is a milestone work on co-word analysis. The first part of the book is an introduction on how to study the force of science. The second part is an analysis of the power of texts in science and technology, in which the authors have presented the theoretical foundation of co-word analysis, that is, "actor network." The third part is a detailed description of co-word analysis with examples. The last part is a conclusion.

Since publishing this book, co-word analysis has spread to researchers from not only France, but also the United Kingdom, the Netherlands, the United States, and some other countries. The process, measurement, and interpretation of co-word analysis has been improved to a great extent through these subsequent studies.

Theoretical Foundation—Actor Network

The co-word analysis technique was first proposed to map the dynamics of science. The most feasible way to understand the dynamics of science is to take the force of science in present-day societies into account. "Actor network" is the theoretical foundation for co-word analysis to map the dynamics of science (Callon, 1986).

Laboratories and literatures are considered as two powerful tools for scientists to change the world—they build complex worlds in laboratories and enforce them on paper (Latour, 1987). This implies that scientists attach particular importance to texts. They are not only using texts to publish their world built in the lab but also using texts as a way to build a world and enroll others. Even though science cannot be reduced to texts only, texts are still a prime source for studies on how worlds are created and transformed in the laboratory. Therefore, instead of following the

actors to see how they change the world, following the texts is another way to map the dynamics of science.

Based on the co-occurrence of pairs of words, co-word analysis seeks to extract the themes of science and detect the linkages among these themes directly from the subject content of texts. It does not rely on any a priori definition of research themes in science. This enables us to follow actors objectively and detect the dynamics of science without reducing them to the extremes of either internalism or externalism (Callon et al., 1986b).

Overall, co-word analysis considers the dynamics of science as a result of actor strategies. Changes in the content of a subject area are the combined effect of a large number of individual strategies. This technique should allow us in principle to identify the actors and explain the global dynamic (Callon et al., 1991).

Inclusion Index, Proximity Index, and Equivalence Coefficient

The first step of co-word analysis involves extracting keywords from records in indexing databases. After keywords are extracted from each document, a co-occurrence matrix of keywords can be constructed. Analyzing the interesting features of the co-occurrence matrix is the final and most important step of co-word analysis.

As different questions may be asked about the network of science, the co-occurrence matrix is subjected to various operations. A general co-word analysis is focused on two of these questions: one is to detect the hierarchies among the areas of a research problem, and the other is to detect the minor but potentially growing areas. In the early studies of co-word analysis, two indexes were introduced to address these two questions (Callon et al., 1986c).

The hierarchies of subject areas in a research problem can be detected by calculating an index, called the inclusion index (I_{ij}):

$$\mathbf{I_{ij} = C_{ij} / \min (C_i, C_j)} \qquad (1)$$

where,

C_{ij} is the number of documents in which the keyword pair (M_i and M_j) appears;
C_i is the occurrence frequency of keyword M_i in the set of articles;
C_j is the occurrence frequency of keyword M_j in the set of articles;
min (C_i, C_j) is the minimum of the two frequencies C_i and C_j.

I_{ij} has a value between 0 and 1, and it can be interpreted as a conditional probability. When $C_i > C_j$, that is, M_i is more general than M_j and includes M_j sometimes, I_{ij} measures the probability of finding M_i in an article given that M_j appears in it. An extreme case is that when $I_{ij} = 1$, M_j is fully included by M_i, that is, the M_j always co-occurs with M_i in the same article. The probability of finding M_i is 1, given M_j is found in the same article.

However, sometimes, even though I_{ij} has a low value, it is still significantly greater than the unconditional probability of finding M_i in any one of the N articles in the collection. Such a situation implies that there are some mediator keywords, which have a relatively low occurrence frequency but still have significant relationships with some of the peripheral keywords. To bring out such patterns, a proximity index P_{ij} is defined:

$$\mathbf{P_{ij} = (C_{ij} / C_i C_j) \bullet N} \qquad \textbf{(2)}$$

C_i, C_j, and C_{ij} have the same meaning as in formula (1). N is the number of articles in the collection. The mediator and peripheral keywords pulled out by P_{ij} represent minor but potentially growing areas.

In later co-word studies (e.g., Turner et al., 1988; Whittaker, 1989; Law & Whittaker, 1992; Coulter et al., 1996; Coulter et al., 1998), another index is employed to calculate the association values between word pairs. This coefficient is called the equivalence index (e-coefficient) (Callon et al., 1991) or strength (Coulter et al., 1998). It is defined as follows, where C_i, C_j, and C_{ij} have the same meanings as in formula (1):

$$\mathbf{E_{ij} = (C_{ij} / C_i) \bullet (C_{ij} / C_j) = (C_{ij})^2 / (C_i \bullet C_j)} \qquad \textbf{(3)}$$

E_{ij} has a value between 0 and 1. Similar to (1), E_{ij} measures the probability of word i appearing simultaneously in a document set indexed by word j and, inversely, the probability of word j if word i appears, given the respective collection frequencies of the two words. For this reason, E_{ij} is called "a coefficient of mutual inclusion" by Turner and his colleagues (Turner et al., 1988).

Inclusion Map, Proximity Map, and Sub-Networks

After the inclusion and proximity indexes are calculated, inclusion and proximity maps are created. The inclusion maps are designed to discover the central themes in a domain and depict their relationship to keywords that occur less frequently. The proximity maps are designed to discover connections between minor ideas hidden behind the central themes. These two kinds of maps correspond to two general types of studies. The first type of study involves getting more information about a certain theme. The second category of study concerns the analysis of the links between themes.

To create inclusion maps, the link that has the highest inclusion index value is selected first. These linked nodes become the starting points for the first inclusion map (subnetwork). Other links and their corresponding nodes are then added into the map in the decreasing order of their inclusion index until the threshold I_0 is reached. All nodes contained in the resulting cluster are removed from consideration as candidates in subsequent maps. The next map then starts with the link of highest inclusion index value of the remaining links. Keywords that appear on

the top level of inclusion maps are called "central poles" of the domain of research. Keywords that are included in the central poles, and themselves include some other words at lower levels, are called "mediator words" (Callon et al., 1986c).

The process to create proximity maps is similar to that for inclusion maps. The difference is that the proximity index is used instead of the inclusion index. If the threshold P_0 is lowered enough, more proximity connections between keywords will appear in the map and, eventually, the mediators and central poles found in inclusion maps will reappear. In this way, the relationship between minor issues and central poles can be studied (Callon et al., 1986c).

There is another method to construct clusters (or subnetworks) consisting of keywords that are more strongly linked internally than with keywords external to this sub-network (Callon et al., 1991). Essentially, this is similar to the inclusion maps above. The clusters could correspond to centers of interest in the research problem that are intensively studied by researchers. However, instead of using the inclusion index and threshold I_0, an e-coefficient is used in this method to measure the strength between keywords, and a threshold of ten is used to limit the number of words in one subnetwork. The procedure still starts from the link with the highest e-coefficient. When a cluster already has 10 words in it, the next link will be refused. The value of this link that is first refused is called the saturation threshold. After a cluster saturates, a new cluster is started. The e-coefficient value of the first link of this new cluster is called the "ceiling threshold." Based on the association value of the inter-cluster link and external links and the value of the ceiling threshold and saturation threshold, three distinct categories of clusters can be identified. The first category is isolated clusters, which are characterized by an absence (or low intensity) of links with other clusters. The second is secondary clusters, whose external links with other clusters above the ceiling threshold are sufficiently strong that it is legitimate to consider that they are the natural extension of one of these. The third is principal clusters, to which one or more other (secondary) clusters are associated by links whose value is lower than the saturation threshold.

Coulter et al. (1998) have divided the process of constructing subnetworks into two "passes." During Pass-1, the network is constructed similar to the process of creating inclusion maps above, but the e-coefficient is used to measure the strength of association between two keywords. In Pass-2, the network is extended by adding Pass-2 links. To be a Pass-2 link, both nodes of the link must be included in some Pass-1 network.

Density, Centrality, and Strategic Diagram

An earlier study was carried out to compare citation, co-citation, and co-word analyses of the state of five disciplines (Healey et al., 1986). It was

found difficult to analyze and accept the preliminary co-word results, and some experts doubted the reliability of the findings. The co-word technique evaluated in this study is called the "first generation" of co-word analysis by Law et al. (1988). A "second generation" analysis is presented in the same article to overcome the problems encountered in the comparison study.

In the "second generation" co-word analysis, a strategic diagram is used to illustrate the "local" and "global" contexts of research themes. This diagram is created by putting the strength of global context on the *X* axis and putting the strength of local context on the *Y* axis. This diagram is used in many later co-word studies. Two kinds of indexes (i.e., density and centrality) are used to measure the strength of local context and global context respectively.

Density. Density is used to measure the strength of the links that tie together the words making up the cluster; that is the internal strength of a cluster. It provides a good representation of the cluster's capacity to maintain itself and to develop over the course of time in the field under consideration (Callon et al., 1991). Ranking subject areas (clusters) in terms of their internal coherence (density) is designed to provide information for systematic discussion of a major policy alternative. Further, sorting the keywords by decreasing order of density can provide a precise description of the areas (Bauin et al., 1991).

The value of the density of a given cluster can be measured in several ways. Generally, the index value for links between each word pair is calculated first. Then, the density value can be the average value (mean) of internal links (e.g., Turner et al., 1988; Coulter et al., 1998), the median value of internal links (e.g., Courtial et al., 1993), or the sum of the squares of the value of internal links (e.g., Bauin et al., 1991). An internal link means both of the words linked by it are within the cluster.

Centrality. Centrality is used to measure the strength of a subject area's interaction with other subject areas. Ranking subject areas (clusters) with respect to their centrality shows the extent to which each area is central within a global research network. The greater the number and strength of a subject area's connections with other subject areas, the more central this subject area will be in the research network (Bauin et al., 1991).

For a given cluster (area), its centrality can be the sum of all external link values (e.g., Turner et al., 1988; Courtial et al., 1993) or the square root of the sum of the squares of all external link values (e.g., Coulter et al., 1998). More simply, it can be the mean of the values of the first six external links (e.g., Callon et al., 1991). An external link is a link that goes from a word belonging to a cluster to a word external to the cluster.

Strategic Diagram. A strategic diagram that offers a global representation of the structure of any field or subfield can be created by plotting centrality

and density into a two-dimensional diagram (Law et al., 1988). Typically, the horizontal axis represents centrality, the vertical axis represents density, and the origin of the graph is at the median of the respective axis values. This map situates each subject area within a two-dimensional space divided into four quadrants.

The strategic diagram is used in many co-word analysis studies (e.g., Turner et al., 1988; Courtial & Law, 1989; Turner & Rojouan, 1991; Callon et al., 1991; Coulter et al., 1998) and the analysis based on it is similar among these studies. Generally, the subject areas in quadrant 1 are both internally coherent and central to the research network in question. However, those areas in quadrant 4 seem to be of only marginal interest to work in the global research network. Coherent subject-specific areas always appear in quadrant 3 of the diagram. These areas are internally well structured and indicate that a constituted social group is active in them. However, they appear to be rather peripheral to the work being carried out in the global research network. Weakly structured areas are found in quadrant 2. These subjects, individually, are linked strongly to specific research interests throughout the network but are only weakly linked together. In other words, work in these areas appears to be underdeveloped, but it could potentially be of considerable significance to the entire research network. All these characteristics of a strategic diagram can be summarized in Figure 5.

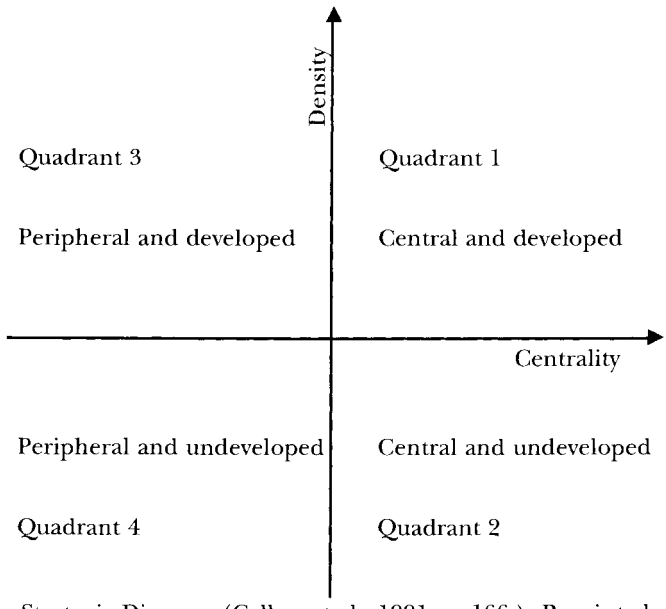


Figure 5. Strategic Diagram (Callon et al., 1991, p. 166.) Reprinted with kind permission from Akadémiaí Kiado publishers from an article in *Scientometrics*, 22(1), 155-205.

Comparative Analysis of Networks

The Stability of Networks. A striking feature of some strategic diagrams is the radical change in the configuration of the research network at two periods. This reflects the dynamics of science. Based on the strategic diagram, we can analyze the stability of the networks and foresee their changes in the future. This issue is addressed in many studies, and the methods used in these studies fall into two categories.

The first method used to study the stability of networks is directly based on the strategic diagrams (e.g., Callon et al., 1991; Turner & Rojouan, 1991). The findings can be summarized as showing that the probability for the research content of themes situated in quadrants 2 and 3 to change over time is significantly higher than it is for themes which are situated in quadrant 1. With a low density, the unstructured themes in quadrant 2 tend to undergo an internal structuring to improve their cohesiveness. With a low centrality, the scope of themes in quadrant 3 is likely to be extended in order to better articulate what is being done in the rest of the network. The reason as well as the goal for all these changes is to situate their work at the heart of their research network (quadrant 1). This can be done either by enlarging its scope or by improving its visibility through conceptual developments in the definition of a research program.

The second method is based on the ratio of centrality to density (c/d) (e.g., Courtial et al., 1993; Turner et al., 1994). The ratio (c/d) is considered as a meaningful indicator of the development stage of science and technology by many researchers. On the one hand, the findings show that, if this ratio tends toward 1, it indicates that this area is serving as a mainstream in the research network and is capable of redefining the global configuration of the system. On the other hand, if this ratio tends away from 1, it indicates the theme is falling out of favor and could well disappear as a subject of interest in the research network. However, Leydesdorff (1992a) claims "the c/d ratio is indeed a measure of the mutual information provided between the word distribution and the document distribution in that part of the structure" (p. 310) and cannot be used for this purpose.

Network Comparison. In co-word analysis studies, several subnetworks can be constructed concurrently while each network changes over time. To detect the difference among subnetworks simultaneously or subnetworks at different times is another issue studied by many researchers.

The comparison of two networks, N_1 and N_2, which might be two networks at different times or two distinct networks at the same time, can be done by a three-stage method (Callon et al., 1991).

The first stage is to compare the clusters. Let C_{1i} be the set of clusters of network N_1 and C_{2i} be the set of clusters of network N_2. A transformation

index (also called a dissimilarity index) is defined to measure the degree of dissimilarity between two given clusters. This index is defined as:

$$t = (W_i + W_j) / W_{ij} \quad (4)$$

where,

W_i is the number of words in cluster C_i;
W_j is the number of words in cluster C_j; and
W_{ij} is the number of words common to C_i and C_j.

For example, if the cluster C_i is defined by seven words and the cluster C_j by four words, and if four words among these eleven words are common to the two clusters, the transformation index is $t = 11 / 4 = 2.8$.

The second stage is to compare the positions in the strategic diagrams of those clusters demonstrated to be similar in stage one. This comparison can go beyond a simple enumeration of correspondences between clusters and bring out the relative position and degree of development of similar clusters within their respective networks.

The third stage is to create the life cycle curve of clusters in the case of dynamic analysis for a research network at times T_0, T_1, T_2 ... T_{10}. Suppose a set of similar clusters is identified in a comparative analysis at different times where C_{10} from T_0 corresponds to C_{11} from T_2; C_{11} corresponds to C_8 from T_3; and so on. This set of similar clusters is called a series. Clearly, the more stable a network is, the more series there are indicating the temporal propagation of its clusters. The existence of these series provides information about the progressive transformation of the clusters through time.

The transformation of networks and their intersections with other networks across time periods provides insights into the emergence of research themes. The similarity of networks in different time periods is also studied by Coulter et al. (1998). In this study, the authors employ the similarity index (SI), which comes from Callon's dissimilarity (or transformation) index above. It is defined as follows:

$$SI = 2 \bullet (W_{ij} / (W_i + W_j)) \quad (5)$$

where,

W_i is the number of descriptors in network N_i;
W_j is the number of descriptors in network N_j; and
W_{ij} is the number of descriptors common to N_i and N_j.

A constant 2 is multiplied to make the maximum value of SI to 1, which occurs when N_i and N_j have identical nodes. SI is used to measure the intersection of the descriptors in two networks and to examine the emergence of a network during a particular period.

Index of Influence and Provenance. Another comparative analysis is done by Law and Whittaker (1992) to highlight the overlap between themes on

similar subjects in succeeding time periods. Two indexes, the Index of Influence (I) and the Index of Provenance (P), are employed to measure the degree of continuity between themes in generations. These two indexes are calculated as follows:

$$\mathbf{I_{ij} = (2 \bullet M_{ij} + Ln_{ij}) \ / \ (2 \bullet N_i);} \qquad \textbf{(6)}$$

where,

M_{ij} is the number of words in both theme i and (succeeding) theme j;
Ln_{ij} is the number of words in both theme i and linked to subsequent theme j but belonging to no other theme in this generation; and
N_i is the number of words in theme i.

$$\mathbf{P_{ij} = (2 \bullet M_{ij} + Ln_{ij}) / N_j;} \qquad \textbf{(7)}$$

where,

M_{ij} is the number of words in both theme j and (preceding) theme i;
Ln_{ij} is the number of words in both theme j and linked to preceding theme i but belonging to no other theme in this generation; and
N_j is the number of words in theme j.

The index I_{ij} shows the proportion of the words within a theme in one generation attached to any given theme in the next generation. A high I_{ij} means that the "influence" of a first generation theme on one of the second generation is high. The P_{ij} index shows the proportion of words within a second generation theme that come from any given theme in the preceding generation. A high P_{ij} means that the "provenance" of a second generation theme primarily lies in a single theme of the first generation. Using these two indexes, the authors analyzed the continuities between themes and identified the lines of work in the field of acidification research. They were satisfied that they had detected a number of relatively stable themes by means of I and P indexes.

Frequency Analysis, Proximity Analysis, and Database Tomography

Database Tomography is a patented system for analyzing large amounts of textual computerized material (Kostoff et al., 1995). It can be considered as another generation of co-word analysis. Algorithms for extracting multi-word phrase frequencies and performing phrase proximity analysis are included in this system. Phrase frequency analysis can be used to discover the pervasive themes of a database while the phrase proximity analysis can be used to detect the relationships among these themes and between these themes and subthemes (Kostoff et al., 1997a). The indexes used in Database Tomography are similar to those used by traditional co-word analysis, such as the E_{ij} equivalence index. But the co-occurrence of keywords is limited to ± 50 words within the text.

Similar to co-word analysis, Database Tomography can identify the main intellectual thrust areas and the relationship among these thrust areas. It provides a comprehensive overview of a research network and allows specific starting points to be chosen rationally for more detailed investigations into a topic of interest. Kostoff and his colleagues have employed Database Tomography tools to study chemical literature (Kostoff et al., 1997a). There are two appendixes in their article that show Database Tomography can be used for the generation of taxonomies and the identification of promising research directions.

Based on the term co-occurrence information, Database Tomography can also be used to expand the initial query in information retrieval (IR) systems and, in turn, allow the retrieval of relevant documents that would not have been retrieved with the initial query (Kostoff et al., 1997b). Simulated nucleation is the name given to the form of Database Tomography adapted to IR. In simulated nucleation, a core nucleus is developed first, and similar material is added as time develops until the desired amount of material is obtained. Then the main algorithms of Database Tomography (phrase frequency and phrase proximity analysis) operate on this core group of documents to identify patterns of word combinations in existing fields and generate new search term combinations that follow the newly identified patterns. The process is repeated until convergence is obtained, where relatively few new documents are found even though new search terms are added. Thus, simulated nucleation is running in a self-correcting cybernetic homeostatic model. It continually expands the coverage and improves the quality of the retrieval results.

Rotto and Morgan (1997) have employed frequency analysis and phrase proximity analysis techniques to study if the work in a dissertation abstract is potentially applicable to industry. The study first counts the frequency of every single-, double-, and triple-word phrase. Then, using the most frequently occurring technology-related word phrases as theme words, phrase proximity analysis is applied to construct clusters of word phrases that co-occur within abstracts. These clusters are then examined to investigate whether research subspecialties or related research focuses could be identified.

The Advantages and Limitations of Co-Word Analysis

Advantages of Co-Word Analysis

Quantitative Over Qualitative. The drawbacks of qualitative methods have already been addressed at the beginning of this discussion. The advantages of co-word analysis over qualitative analysis were recognized by researchers from the time of its introduction. In the book by Callon et al. (1986a), the advantages of co-word analysis over qualitative methods have been shown at several points, for example:

> The problem of distinguishing between the successful and the less successful strategies of translation in qualitative analysis is solved by quantitative means: the aggregation of the co-occurrences of signal-words across a population of texts and the depiction of significant levels of such co-occurrences by graphical methods . . . . Using the quantitative in pursuit of the qualitative, we are also able to highlight features of scientific fields that have not always been recognized. The heterogeneity of scientific world-building is preserved in co-word analysis, where experimental findings, research methods, concepts, social problems, material artifacts and locations may appear together on the maps. (Callon et al., 1986d, pp. 225-26)

Qualitativists often jump from detailed analyses of scientific controversy to general explanations posed in terms of social interests:

> [Qualititativists] are unable to make the connection in a more detailed and less perilous manner. By contrast, the co-word approach, by summarizing articles in terms of forceful words and counting occurrences and co-occurrences to trace developments at an aggregate level, not only allows successful translations to be traced and distinguished from those that quickly disappear; it also makes it possible to uncover the many direct and indirect links that exist between translations whether or not these lead rapidly to social problems and interests. (Callon et al., 1986c, p. 108)

Flexibility. Compared with other methods of analysis that focus on texts, co-word analysis is much more flexible in that it shows the research network with graphs. On the one hand, these graphs can be simplified to show the overall structure of the network. On the other hand, one can zoom in on certain areas and trace the co-word patterns in as much detail as one wishes.

When Callon (1986) studied a collection of patents using co-word analysis, he took the flexibility advantage of co-word analysis and applied two techniques to analyze the maps. The first technique was to simplify the maps. It was found that certain words unified the whole of the field without really adding new information. When they were deleted, the structure of the maps was simplified but not altered. The second technique was to zoom in on a pole. The author used zooms as a means to carry out a detailed study of why a concept (i.e., "enzyme") totally disappeared from the inclusion maps in 1981 after having been a central pole in 1980. This zooming technique showed that this abrupt change between the two periods was more than a simple fluctuation; it was linked to the appearance of a small number of patents that introduced new centers of interest and reorganized existing relationships.

The technique of zooming in on certain areas to get more information on a specific word of medium frequency has also been used in other studies (e.g., Callon et al., 1986d). In addition, there is a technique proposed to do variable level clustering, which is another flexible way to show the maps of research areas at different levels (Turner et al., 1988).

Limitations of Co-Word Analysis

It is obvious that the quality of results from co-word analysis depends on a variety of factors, such as the quality of keywords and index terms, the scope of the database, and the adequacy of statistical methods for simplifying and representing the findings (Law et al., 1988). Solely making use of keywords and index words was the biggest problem of early co-word analysis. It was called "indexer effect" and was addressed by many researchers.

Callon et al. (1986d) mentioned one such problem when dependence was on indexing:

> Indexing is an intervention between the text and the co-word analysis, and the validity of the map will depend, to a certain extent, on the nature of the indexing. Yet since indexers try to capture what it is about a text that is interesting, they partially reproduce the readings that the texts are given within the field itself. Thus, despite the fact that indexing is not entirely reliable, validity is never totally absent. (p. 226)

Turner et al. (1988) questioned the schemes used in co-word analysis as follows:

> However, most of the work done in this area has used the classification schemes of the data base producers to draw conclusions. Designed for document retrieval, these schemes are generally not suited for monitoring changes in the state of technological art at any given moment in time. (pp. 320-21)

Whittaker (1989) has pointed out that the results of co-word analysis are dependent on how the indexers choose the keywords to conceptualize the scientific fields. However, the results from indexing are more akin to the conceptualizations of indexers than to those of the scientists whose work is being studied.

In addition to the "indexer effect," some other limitations of co-word analysis are also recognized by researchers, which include, but are not limited to, the following:

1) The representations of the results given by co-word analysis are too difficult to read. (Whittaker, 1989)
2) The coverage of database is incomplete. Certain types of literature, such as patents and the"grey literature," lie outside of the publication circuit and are not indexed in the database. This means that the results from co-word analysis cannot reflect the whole picture of the research field in question.
3) The delay between the writing of a document and the moment when it is indexed and entered into a database causes the co-word analysis to fail to detect emerging research themes at an early stage. (Callon et al., 1986)

Some Issues in Co-Word Analysis

Assumptions

The assumptions of co-word analysis are presented by Whittaker (1989) as follows:

> It [co-word analysis] relies upon the argument that (i) authors of scientific articles choose their technical terms carefully; (ii) when different terms are used in the same articles it is therefore because the author is either recognizing or postulating some non-trivial relationship between their referents; and (iii) if enough different authors appear to recognize the same relationship, then that relationship may be assumed to have some significance within the area of science concerned. (p. 473)

A fourth premise, that the keywords chosen by trained indexers as descriptors of the contents of articles are in fact a reliable indication of the scientific concepts referenced in them, makes it possible to use the keywords as the basic data for co-word analysis.

Later, Law and Whittaker (1992) have restated two of the assumptions above. First, co-word analysis assumes that the keywords used by indexers to index a paper reflect the present stages of the scientific research in question. Then, co-word analysis assumes that arguments received by other scientists will lead to the publication of further scientific papers that are indexed by similar sets of keywords.

If all these are reasonable assumptions, it is then possible for co-word analysis to make use of the frequencies of word pairs in an article set as a way to map the structure of concepts embodied in the articles.

Indexer Effect

As noted before, the indexer effect is one serious problem with co-word analysis. The main criticism against co-word analysis is also because of this. Many researchers tried to address this issue, and some tests have been done to overcome this problem.

One result of the indexer effect is that keywords assigned to the articles by the indexers are out of date. There are three sources contributing to this problem: (1) the lexicon used by indexers is itself out of date; (2) the indexers may use combinations of keywords that reflect the conventional views of science as they were previously; and (3) the inevitable delay between the publication of an article and the appearance of an entry in the database causes a problem (Whittaker, 1989).

In Law and Whittaker's study (1992), some experts are asked to evaluate the keywords assigned by indexers in the PASCAL database. Even though most comments are positive, three kinds of complaints are posed. First, some keywords assigned by indexers are too general. Second, one or two specific terms had been omitted from the satisfactory list. Third, errors and misplaced specificity are found—i.e., the indexer puts the wrong emphasis, or even a mistaken emphasis, in keywording.

Some tests have been carried out to overcome the indexer effect of co-word analysis. Generally, these tests make use of some mechanism to automatically index the database. Two examples are as follows:

- *Test with QUESTEL-PLUS.* QUESTEL-PLUS is a full text information retrieval system used by TELESYSTEMES in France. In collaboration with TELESYSTEMES, Callon and his colleagues (Callon et al., 1986) combined different techniques with QUESTEL-PLUS and ran LEXIMAPPE together with them. This established a completely computerized chain of procedure running from a QUESTEL-PLUS treatment of full-text literature to the automatic establishment of inclusion or proximity maps by the LEXIMAPPE.

  The study has tested the chain on a small dietary fiber file. In comparison with the manually indexed file, the results obtained are encouraging in three aspects. First, the general and redundant words, which complicate the maps without adding new information, are eliminated. Second, a much larger number of specific peripheral issues appear in the inclusion maps. Third, the structure of the proximity maps is much richer and more detailed.
- *Test with LEXINET.* Turner et al. (1988) have carried out a test to overcome the limits of manual indexing through the use of a computer-assisted indexing system known as LEXINET. The goal of the LEXINET system is to help an expert construct an indexing vocabulary suitable for a particular area of study by an interactive validation process between the expert and the system.

  The study shows that, with LEXINET, the indexing process is considerably accelerated. Since part of the delay between the writing of a document and the moment it is available for analysis is caused by manual indexing, using LEXINET can reduce the time lag considerably. Consequently, it improves the quality of the information available for a co-word analysis and essentially reduces the indexer effect.

During the last two decades, much progress has been made in the field of automatic indexing. With the development of automatic indexing, we should be able to considerably reduce, if not eliminate, the "indexer effect" found in the results of co-word analysis.

Related Statistical Methods

The statistical method used in co-word analysis is similar to the single linkage cluster algorithm. This method is simple and considered unreliable. Some other statistical methods have been studied to consider the possibility of using them in co-word analysis.

Courtial (1986) has compared correspondence analysis and multidimensional scaling with co-word analysis and indicated the limitation of the first two methods as follows:

- Since the goal of correspondence analysis is to extract a set of dimensions of decreasing importance in the same way as principal component analysis does for quantitative characteristics, the representation of objects or characteristics is limited to the space created by the first two dimensions. Applying this method in a test, the two first dimensions merely "explain" 11.2 percent of the total distances between keywords. The reason is that keyword coded scientific articles never have the usual features of characteristics attributed to objects. There is an inherent difference between keywords and characteristics. Keywords cannot be treated as characteristics if the associations between these characteristics must be the combinations of a small number of independent dimensions.
- Multidimensional scaling suffers from the same sort of difficulties. The goal of this method is to identify a configuration of words such that the calculated distances between the words can reflect the geometric distances as much as possible. This is done within a space, which is set at two or three dimensions beforehand. When applying this method in a test, it is possible to find some global properties of the field. But the results do not allow any more detailed analysis because the stress is far from negligible (pp. 190-92).

Leydesdorff has employed factor and clustering analysis techniques in his co-word analysis. The approach is described as follows:

> Co-word analysis generates a symmetrical matrix with an empty diagonal, i.e. word A AND word B happens as many times as B AND A. The matrices are factor-analyzed using both orthogonal and oblique rotations (to check for inter-factorial relations). For graphic representation, cluster analysis was pursued using Wards' mode of analysis with the cosine as the similarity coefficient. (as cited in Whittaker, 1989, p. 489)

There is an important difference between Leydesdorff's co-word analysis and the co-word analysis we have described here. The former uses some complicated statistical techniques to assign words into clusters while the latter does not. The latter rests more upon the assumption that there is a cluster-type structure and its algorithm is set to build those clusters link by link according to the relative frequencies of words and co-words in the document. The goal of the former method is to identify, list, and measure the distance between classes to create distinction rather than emphasizing connection and continuity. By contrast, the goal of the latter is to describe a network of words and explore the qualitative character of the links between them by concentrating on, and tracing out, connections and crossroads in that network. So, the two methods are actually doing different jobs and are appropriate for different purposes. Whittaker (1989) compared these two methods in his study. He thinks that, if one is dealing

with a relatively homogeneous set of documents, it may be reasonable to assume that all the nontrivial title words should be included in the cluster structure. If the task is more complicated and the analysis is on a large and heterogeneous set, such an assumption seems unwarranted, and the method we have described here offers significant advantages.

Turner et al. (1994) have studied the co-word analysis techniques in connection with the local components analysis (LCA) in the GEODE (La Gestion Optimisee des Documents Electroniques) project. LCA is a neural algorithm designed to identify "data poles" and their "influence areas" in a document set. It can reveal local data structures in a very large data set. In this study, LCA was used to produce a data pole map. Each data pole is a constructed object in the GEODE system and described in the same way as the LEXIMAPPE generated objects—i.e., by a list of keywords. However, the LCA technique can supply additional information: it ranks the documents in a data flow according to their contribution to defining the emergence of a data pole.

Nederhof and van Wijk (1997) analyzed a co-occurrence matrix of the 104 most frequent nontrivial topics and 63 SSCI journal groups. They computed and transformed a discipline by topic correlation matrix into a discipline by discipline matrix. Disciplines with high correlation (Pearson $r > 0.88$) were merged. Two data sets were analyzed in this study. One set consisted of topics on which publication changed greatly, and that gives rise to a "dynamic" map. The other set consisted of a matrix of about 100 nontrivial topics that most frequently occurred in SSCI in 1986-1990, through which the "static" map was generated. Both sets of matrices are analyzed by means of combined cluster analysis and correspondence analysis. Both topics and disciplines were clustered separately but analyzed jointly in the correspondence analysis. Compared to co-word methods, this set mapping method has the important advantage that it related not just words to words, but also, in one single map, disciplines to disciplines and topics to disciplines.

Measurements

In addition to the measurements described in previous sections, researchers have also studied the probability of making use of some other measurements.

The usability of the Jaccard index and "statistical coefficient" was studied by Courtial (1986) as follows:

- The Jaccard index is often used to express the degree of intersection between two document sets and it is defined as:

$$J_{ij} = (C_{ij}) / (C_i + C_j - C_{ij}) \qquad (8)$$

This index can be used to measure the relative degree of overlap between "semantic areas" of words within a given database. However, it

cannot handle associations between low-frequency and high-frequency words very well, because it will have low values even when the low-frequency word always appears together with the high-frequency word. Therefore, the author thinks, this index can only be used to explore overlap between medium-frequency words.

- "Statistical coefficient" is similar to the proximity coefficient. It can be used to compare the observed frequency (C_{ij} / N) of a pair of words with the expected frequency of that pair if the words were independent (C_i / N • C_j/ N). Compared with the proximity coefficient, this coefficient has the advantage of being symmetrical and normalized. It is calculated as:

$$S_{ij} = 1 / S \bullet (C_{ij} - C_i C_j / N) \qquad (9)$$

where S is the standard deviation of the hypergeometrical distribution. According to Courtial, this coefficient is not usually used because the strength of association is not an important variable in the graphs. In addition, the computation of this coefficient takes a long time, while the extra information is not essential for interpretation.

In the study of Coulter et al. (1998), co-word analysis has been used to get an evolutionary perspective of software engineering. In order to measure the complexity of networks, they use the ratio of links to nodes L/N as a measurement. As (N-1)/N L/N (N-1)/2, the minimum value for L/N is ½. "Percentage of connectivity" is another related and normalized measure of a network's complexity. It is based on the ratio of the number of links in a network to its maximum possible number of links, that is, 2L/(N(N-1)). This value will be greater for simple stand-alone networks or subnetworks than that of larger networks because the numbers of nodes and links are fewer.

How to Interpret the Map

The maps obtained by co-word analysis are generally considered very difficult to understand by themselves. They have to be interpreted with caution. It is suggested by Callon et al. (1986) that the interpretation must be active and based on the comparison of inclusion and proximity maps. In some cases, it is necessary to make use of zooms and examine the original documents (or at least their descriptors). Collaborating with the experts is another way to improve the interpretation.

As the goal of co-word analysis is not to photograph a field of knowledge but to reveal the strategies by which actors mutually define one another, Callon and his colleagues (1991) suggest that the maps cannot be considered statistically and they must be interpreted dynamically. Attention should be paid to not only the internal dynamics of each network but also the interactions between networks. For the internal dynamics, we need to analyze the appearances, disappearances, transformations, and

movements in the series of clusters and the overall life cycles of clusters. For the interactions between networks, possible interactions include academic networks to general network, applied research to academic research, and some other complex interactions.

Where Should the Words Come from?

In 1987, Leydesdorff criticized the co-word analysis technique for the indexer effect, and his answer to this problem was to use title words instead of keywords as the basis of co-word analysis (as cited in Whittaker, 1989). This idea looks attractive because it might allow more direct access to the views of authors, and the descriptions can give more confidence to those who have doubts about the indexing process.

However, Whittaker (1989) points out that there are two difficulties in using title words. One is that authors might choose their title words deliberately in order to address a particular readership and produce an "audience effect." The other concerns the usage of nonstandard titles such as those in the form of a rhetorical question. To discover whether title words are preferable to keywords for co-word analysis, Whittaker has carried out a comparative study. He found that keyword analysis generates a picture similar to, but substantially more detailed than, that created by title word analysis. It does not show that either form of analysis is superior to the other. To some extent, this also proves the indexer effect is not a problem, at least in this case.

A literature review shows that the words used in co-word analysis are expanding from keywords in a lexicon to words in the full-text. In early studies, only keywords from a lexicon are used (e.g., Bauin, 1986). Later, the documents are indexed by title, summary, and a certain number of restricted keywords (descriptors) drawn from a lexicon from the study by Callon et al. (1991). Recently, Rotto and Morgan (1997) suggested co-word analysis could be performed on abstracts using words suggested by industry experts to help identify more specific research focuses within the research area of need. Finally, in Database Tomography, full-text words are used. One of the many advantages of full text over key or index words is the ability to retain low frequency but highly important phrases, since the keyword approach ignores the low frequency phrases (Kostoff et al., 1997a).

To What Extent Should the Words be Normalized?

Even after we have decided where to get the words, we still need to "normalize" them before we do a co-word analysis. Normalization has been addressed and done in several previous studies.

Turner et al. (1988) note that databases for information retrieval generally have to be "cleaned up" when they are used as a science evaluation tool. Strategies have to be devised to normalize institutional addresses, and country and author names, overcome the limits of manual indexing, and deal with multi-authored papers.

In the study by Courtial et al. (1993), the normalized title is used as a list of keywords. The WPIL patent database used in their study provides a normalized title for each patent family of the database, which is given by WPIL editors. These improved titles are based on the whole text of the priority documents. In addition, WPIL also makes use of thesaurus terms. Word processing is even used to improve the list of uniterms by joining a set of two succeeding words, such as joining "ice cream" to make "ice*cream." All these pre-processes on keywords enable the authors to obtain meaningful results in the study.

Nederhof and van Wijk (1997) have studied the association among topics in a discipline. The topics in the study are derived from words in the title of each article. To exclude idiosyncratic terms, only topics occurring at least ten times in a five-year period are analyzed. Many words for which British and American spellings differ have been standardized to the American spelling by the Institute for Scientific Information when they are put into the citation index databases.

Validation of the Method

Validation of LEXIMAPPE. Leydesdorff (1992a) has employed information theory to evaluate the LEXIMAPPE method of co-word analysis of scientific texts. LEXIMAPPE is criticized as follows:

> In LEXIMAPPE, only the strength of the association is computed, the strength of the association is not tested against an expected value for significance in terms of the distribution. As a consequence, two words which most strongly differ in terms of "structural equivalence" may occur in one cluster, and two words which do correlate significantly in terms of their distribution over the document set may occur in different clusters. The basic model is a graph-analytic relational model limited to diadic relations only. (p. 297)

In summary, Leydesdorff has shown that LEXIMAPPE uses rather different mechanisms to cut the "cake": on the basis of the comparisons of distributions, one finds completely different groupings than those abstracted on the basis of single co-occurrences.

In reply to Leydesdorff, Courtial (1992) first questions whether information theory can be used as an evaluation tool. He thinks information theory, when dealing with codes in a universe that are infinite, such as knowledge, confuses equiprobability and information, thus confusing disorder and information. In general, Courtial thinks Leydesdorff's article seems to deplore the attention paid by co-word analysis to infrequent but strong links to the detriment of global statistics.

Leydesdorff (1992b) gives the following reply to Courtial's comments. He thinks the relational algorithm (in LEXIMAPPE) informs us only about how the system reconstructs the information in the data and nothing about what this change means within the network. The relational approach

exhibits relations and hierarchies, not position and dimensions. In order to assess change and continuity, Leydesdorff thinks, one needs a hypothesis with respect to dimension (e.g., in order to know how to assess the author correlation in the data).

Representativeness of Co-Word Analysis. In another article, Leydesdorff (1997) questions co-word analysis again. He thinks words and co-words cannot map the development of science, because words change position not only in terms of the dimensional scheme of "theory," "methods," and "observation results," but also change in meaning from one text to another. By using the distribution of words over the sections, a clear distinction among "theoretical," "observational," and "methodological" terminology can be made in individual articles but not at the level of the set.

Courtial (1998) has given some comments on Leydesdorff's article above. He claims words are not used as linguistic items to mean something in co-word analysis, but as indicators of links between texts, whatever they mean. In co-word analysis, words are chain indexes, allowing one to compute translation networks. What is important for co-word analysis is not the exact meaning or definition of a word, but the fact that this word is linked to word *X* in one case and word *Y* in another case.

Again, Leydesdorff (1998) insists relational indexes cannot warrant inferences about strategic positions and that information calculus provides a useful tool for combining the static and dynamic analysis.

Selection of Method

The appearance of co-word analysis added another choice in the area of bibliometrics and provided another way to discover knowledge in databases. Similar to co-citation analysis, co-word analysis is also a kind of relational study based on the idea that publications should not be considered as discrete units. Instead, each is built upon others (Turner et al., 1988).

Co-citation and co-word analysis are the two most common methods used for constructing the thematic and strategic map of a field. Then which one should be selected?

In the study by Bauin et al. (1991), there were two reasons for choosing the co-word analysis method. The first reason is because they wanted to study the knowledge structure of the field rather than the relationship between researchers. Co-word analysis is based on the scientific content of publications and it serves their purpose directly. The second reason was methodological. They wanted to test the usefulness of co-word analysis in the process of strategic planning to see if it could be used as a tool in science management.

Callon et al. (1991) have shown why co-word is better than co-citation to study interactions between academic and technological research. The reason is the indicators used by co-citation only show the existence of a

link and cannot give any information on the subject or problem area in question. In order to know if it is scientific research or technology that has been the prime mover of an invention or an innovation, it is necessary to return to the documents themselves and to read the contents of the articles and patents identified. As the indicators used in co-word analysis can reflect the subject themselves, it is not necessary to go back to the original documents in all cases.

A review of the previous studies on co-word analysis shows the technique has been employed in the following types of studies:

- Mapping the dynamics of science (Callon et al., 1986a; Courtial & Law, 1989; Coulter et al., 1998).
- Mapping the structure of scientific inquiry (Whittaker, 1989).
- Mapping interaction between basic and technological research (Callon et al., 1991).
- Evaluating input/output relationships in a regional research network (Turner & Rojouan, 1991).

In addition, it is suggested by Callon and his colleagues (Callon et al., 1986) that co-word analysis should also be useful in the documentation field. It can be employed as a means to classify documents in terms of their evolving centers of interest. From this point of view, it should be useful both for retrospective retrieval and the construction and updating of thesauri.

Conclusion

Previous studies have shown co-word analysis to be a powerful tool to discover knowledge in databases. It has been used to detect the themes in a given research area, the relationship between these themes, the extent to which these themes are central to the whole area, and the degree to which these themes are internally structured. In the last twenty years, co-word analysis has been improved in many aspects. The main progress can be found in two fields:

1. *Source of words.* The early tests used the keywords assigned by indexers. Later, words in the title, summary, and abstract are used. Currently, the technical developments in full-text indexing make it possible to use words in full-text to do a co-word analysis. This will reduce the indexer effects greatly.
2. *Measurements.* The measurements used in co-word analysis have improved. The early co-word analysis used the inclusion and proximity indexes. A more general index, e-coefficient, was proposed later. Density and centrality are two other important measures that enable us to draw a strategic diagram.

However, there are still various kinds of problems remaining in the use of this method. One of the problems is the clustering algorithm. The clustering algorithm in current co-word analysis is very simple. Perhaps the other statistical clustering algorithms would work better. Another problem is the measurements. There are many ways to calculate the value of each index or coefficient. Research is needed to determine the relative effectiveness of the approaches. In addition, the procedure to select the files in the test collection, the elimination of "noise" from the data files, and so on also need further study. Improvements in the method will essentially depend on how it is used.

ACKNOWLEDGMENTS

I am grateful to my advisor, Professor Linda C. Smith, for her help in the whole process of writing this paper—from discussions on the ideas in it to repeatedly editing it. I wish to thank Professor Bruce R. Schatz, P. Bryan Heidorn, and David S. Dubin for providing the list of related literature. I would also like to thank Professor F. Wilfrid Lancaster, Jian Qin, M. Jay Norton, and P. Bryan Heidorn for revising and editing this paper.

REFERENCES

Bauin, S. (1986). Aquaculture: A field by bureaucratic fiat. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 124-141). London: The Macmillan Press Ltd.

Bauin, S.; Michelet, B.; Schweighoffer, M. G.; & Vermeulin, P. (1991). Using bibliometrics in strategic analysis: "Understanding chemical reactions" at the CNRS. *Scientometrics, 22*(1), 113-137.

Callon, M. (1986). Pinpointing industrial invention: An exploration of quantitative methods for the analysis of patents. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 163-188). London: The Macmillan Press Ltd.

Callon, M.; Courtial, J-P.; & Turner W. (1986). Future developments. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 211-217). London: The Macmillan Press Ltd.

Callon, M.; Courtial, J-P.; & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155-205.

Callon, M.; Law, J.; & Rip, A. (Eds.). (1986a). *Mapping the dynamics of science and technology: Sociology of science in the real world.* London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986b). How to study the force of science. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 3-15). London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986c). Qualitative scientometrics. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 103-123). London: The Macmillan Press Ltd.

Callon, M.; Law, J.; & Rip, A. (1986d). Putting texts in their place. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 221-230). London: The Macmillan Press Ltd.

Coulter, N.; Monarch, I.; & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science, 49*(13), 1206-1223.

Courtial, J-P. (1986). Technical issues and developments in methodology. In M. Callon, J.

Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 189-210). London: The Macmillan Press Ltd.

Courtial, J-P. (1992). Comments on Leydesdorff's "A Validation Study of LEXIMAPPE." *Scientometrics, 25*(2), 313-316.

Courtial, J-P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science, 49*(1), 98.

Courtial, J-P.; Callon, M.; & Sigogneau, A. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics, 26*(2), 231-242.

Courtial, J-P., & Law, J. (1989). A co-word study of artificial intelligence. *Social Studies of Science* (London), *19*, 301-311.

Frawley, W. J.; Piatetsky-Shapiro, G.; & Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge discovery in databases* (pp. 1-27). Cambridge, MA: AAAI Press.

Healey, P.; Rothman, H.; & Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*, 233-251.

Kostoff, R. N.; Miles, D. L.; Eberhart, H. J. (1995). *System and method for Database Tomography.* U. S. Patent Number 5440481. August 8, 1995.

Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R.; & Pellenbarg, R. (1997a). Database Tomography for technical intelligence: Comparative roadmaps of the research impact assessment literature and the *Journal of the American Chemical Society. Scientiometrics, 40*(1), 103-138.

Kostoff, R. N.; Eberhart, H. J.; & Toothman, D. R. (1997b). Database Tomography for information retrieval. *Journal of Information Science, 23*(4), 301-311.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press.

Law, J.; Bauin, S.; Courtial, J-P.; & Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. *Scientometrics, 14*(3-4), 251-264.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics, 23*(3), 417-461.

Leydesdorff, L. (1992a). A validation study of "LEXIMAPPE." *Scientometrics, 15*(2), 295-312.

Leydesdorff, L. (1992b). A reply to Courtial's comments. *Scientometrics, 15*(2), 317-319.

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the science. *Journal of the American Society for Information Science, 48*(5), 418-427.

Leydesdorff, L. (1998). Reply about using co-words. *Journal of the American Society for Information Science, 49*(1), 98-99.

Nederhof, A. J., & van Wijk E. (1997). Mapping the social and behavioral sciences worldwide: Use of maps in portfolio analysis of national research efforts. *Scientometrics, 40*(2), 237-276.

Price, D. S. (1963). *Little science, big science.* New York: Columbia University Press.

Rotto, E., & Morgan, R. P. (1997). An exploration of expert-based text analysis techniques for assessing industrial relevance in U.S. engineering dissertation abstracts. *Scientometrics, 40*(1), 83-102.

Turner, W. A., & Callon, M. (1986). State intervention in academic and industrial research: The case of macromolecular chemistry in France. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world* (pp. 142-162). London: The Macmillan Press Ltd.

Turner, W. A.; Chartron, G.; Laville, F.; & Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In A. F. J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 291-323). Netherlands: Elsevier Science Publishers.

Turner, W. A., & Rojouan, F. (1991). Evaluating input/output relationships in a regional research network using co-word analysis. *Scientometrics, 22*(1), 139-154.

Turner, W. A.; Lelu, A.; & Goergel, A. (1994). GEODE: Optimizing data flow representation techniques in a network information system. *Scientometrics, 30*(1), 269-281.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science, 19*, 473-496.

Knowledge Discovery in Documents by Extracting Frequent Word Sequences

HELENA AHONEN

ABSTRACT

AS ONE APPROACH TO ADDRESS THE NEW INFORMATION needs caused by the increasing amount of available digital data, the notion of knowledge discovery has been developed. Knowledge discovery methods typically attempt to reveal general patterns and regularities in data instead of specific facts, the kind of information that is hardly possible for any human being to find. In this article, a method for extracting *maximal frequent sequences* in a set of documents is presented. A maximal frequent sequence is a sequence of words that is frequent in the document collection and, moreover, that is not contained in any other longer frequent sequence. A sequence is considered to be frequent if it appears in at least n documents when n is the frequency threshold given. Frequent maximal sequences can be used, for instance, as content descriptors for documents: a document is represented as a set of sequences, which can then be used to discover other regularities in the document collection. As the sequences are frequent, their combination of words is not accidental. Moreover, a sequence has exactly the same form in many documents, providing a possibility to do similarity mappings for information retrieval, hypertext linking, clustering, and discovery of frequent co-occurrences. A set of sequences, particularly the longer ones, as such may also give a concise summary of the topic of the document.

INTRODUCTION

The research field of knowledge discovery in databases (or data mining) has in the last years produced methods for finding patterns and

Helena Ahonen, Wilhelm-Schickard-Institut für Informatik, University of Tübingen, Sand 13, D-72076 Tübingen, Germany
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 160-181

regularities in structured data, mainly in databases. The studies have included efforts to utilize the existing data about, for example, clients, products, and competition. For instance, patterns in client behavior have been extracted. Unstructured data, particularly free running text, place new demands on knowledge discovery methodology. The representations of knowledge discovered are typically sets of frequently co-occurring items, or clusters of items, that seem to behave similarly in some sense. When the data are structured, they are usually easy to define, that is, what are the parts of data—the occurrence or behavior of the data—that are interesting? Regarding unstructured data, however, this is not at all obvious.

When documents are concerned, the words of these documents may appear to be natural item candidates. The more established fields of information retrieval and natural language processing have traditionally concentrated on words and phrases. The phrases may be linguistic phrases, usually noun phrases, or statistical phrases, which are most often frequent noun-noun or adjective-noun pairs. In data mining research, the solution has been to use keywords from a controlled vocabulary (Feldman & Dagan, 1995; Feldman, Dagan, & Klösgen, 1996), names (of people, companies, etc.), or rigid technical terms that usually are noun phrases. However, verb phrases may also carry important hints on acts and processes similar to the following sequences of words:

bank england provided money market assistance
board declared stock split payable april
boost domestic demand

In this article, a method for extracting frequent word sequences from documents is presented. These sequences can be used as items for further knowledge discovery, but they already represent nontrivial and useful knowledge about documents and the entire document collection. Particularly, the method is able to find the maximal frequent word sequences, which are sequences of words that are frequent in the document collection and, moreover, that are not contained in any other longer frequent sequence. A sequence is considered to be frequent if it appears in at least n documents, when n is the frequency threshold given. The specific demands, due to the characteristics of textual data, include the facts that frequent sequences can be very long and that the frequency threshold has to be set rather low to find any interesting sequences.

Frequent maximal sequences can be used, for instance, as content descriptors for documents: a document is represented as a set of sequences, which can then be used to discover other regularities in the document collection. As the sequences are frequent, their combination of words is not accidental, and a sequence has exactly the same form in many documents, giving a possibility to do similarity mappings for information retrieval, hypertext linking, clustering, and discovery of frequent

co-occurrences. A set of sequences, particularly the longer ones, may also give a concise summary of the topic of the document.

In the next section of this discussion, the entire word sequence discovery process is described, including preprocessing the documents, discovery of maximal sequences, assessing the quality of the discovered word sequences and, finally, the usage possibilities of the sequences. The last section presents experiments conducted using a newswire collection.

Extracting Frequent Word Sequences

The core of knowledge discovery are the algorithms that extract regular patterns in the data. In a practical application domain, however, it is essential to see knowledge discovery not just consisting of fast algorithms but as a process that starts from the data as it is available and ends up in the use of the discovered knowledge for some purpose. In the context of this article, the knowledge discovery process contains the following phases:

Preprocessing of the documents
Discovery of word sequences
Ordering the word sequences
Use of the discovered knowledge

The starting point of the process is ordinary running text. The discovery method to be presented in this section has been developed with such a document collection in mind—i.e., the documents are rather brief. Hence, if longer documents are to be used, it might be necessary to use some kind of fragmentation—e.g., consider each paragraph as a document or use some other advanced method for dividing text into fragments (Hearst, 1995; Heinonen, 1998).

Basic Definitions

In order to formulate the approach presented in this article, some terms have to be defined. Assume that there is a document collection that contains a set of documents. Each document can be seen as a sequence of words and word-like characters. Each word has a unique index that identifies the location of the word both in the document and in the document collection. For instance, in the following, portions of three documents can be seen:

(The,70) (Congress,71) (subcommittee,72) (backed,73) (away,74) (from,75) (mandating,76) (specific,77) (retaliation,78) (against,79) (foreign,80) (countries,81) (for,82) (unfair,83) (foreign, 84) (trade,85) (practices,86)

(He,105) (urged,106) (Congress,107) (to,108) (reject,109) (provisions,110) (that,111) (would,112) (mandate,113) (U.S.,114) (retaliation,115) (against,116) (foreign,117) (unfair,118) (trade,119) (practices,120)

(Washington,407) (charged,408) (France,409) (West,410) (Germany,411) (the,412) (U.K.,413) (Spain, 414) (and,415) (the,416) (EC,417) (Commission,418) (with,419) (unfair,420) (practices,421) (on,422) (behalf,423) (of,424) (Airbus,425)

Actually, the documents are seldom processed as such but rather preprocessed before knowledge is extracted. The possibilities and benefits of preprocessing are discussed in detail later in this article.

Usually knowledge discovery methods express the extracted knowledge as some kind of regular pattern appearing in the data. In the current approach, these patterns are represented as *word sequences.* Like the text itself, a word sequence consists of a sequence of words. It is said that a word sequence *occurs* in a document if all the words contained in the word sequence can be found in the document in the same order as within the word sequence. For instance, in the previous sample documents, the word sequence *(retaliation, against, foreign, unfair, trade, practices)* occurs in the first two documents in the locations (78, 79, 80, 83, 85, 86) and (115, 116, 117, 118, 119, 120). The word sequence *(unfair, practices)* occurs in all the documents, namely in locations (83, 86), (118, 120), and (420, 421).

Naturally, a very large number of word sequences can be found in any document, particularly if sequences of all lengths are considered. The set of all sequences, furthermore, does not contain any knowledge that would not be already contained in the text of the documents. On the contrary, any knowledge would be even more difficult to find. Hence, it is important to consider only word sequences that are frequent in the document collection. A word sequence is said to be *frequent* if it occurs in enough documents to equal at least the given *frequency threshold.* For instance, assuming that the frequency threshold is 10, a word sequence is frequent if it occurs in ten or more documents. Note that only one occurrence of a sequence in a document is counted: several occurrences within one document do not make the sequence more frequent. Different definitions in this respect are, of course, possible.

To restrict the number of word sequences further, a *maximal gap* between words in a sequence is given. That is, the original locations of any two consecutive words of the sequence can have only n words between them at the most if n is the maximal gap. Without this restriction, a large number of short frequent sequences, the words of which are located very far away in the text, would be found—e.g., in the previous example, the sequences *(Congress, foreign), (Congress, practices),* and *(against, practices).*

If a word sequence occurs in a document, it is also a *subsequence* of the sequence of words that constitutes the document. In a similar way, a sequence s is a subsequence of any sequence s' if all the words of s occur in s' in the same order as in s.

If some word sequence is frequent, all of its subsequences are frequent. Hence, if there exists a frequent word sequence of length 10,

1,012 frequent word sequences of length 2-9 are returned. Similarly, if *(dow, jones, industrial, average)* is a frequent word sequence in the collection, all the following sequences are found:

dow jones industrial average
dow jones
dow industrial
dow average
jones industrial
jones average
industrial average
dow jones industrial
dow jones average
jones industrial average

Often, however, the longest possible sequences are the most interesting, and their subsequences do not give more information. Hence, a sequence is returned only if it is a *maximal frequent sequence.* A word sequence *s* is a maximal frequent sequence in the document collection if there does not exist any sequence *s'* in the collection such that *s* is a subsequence of *s'* and *s'* is frequent in the collection. That is, a frequent word sequence is maximal if it is not contained in any other frequent word sequence.

Clearly, when a subsequence does not have any independent meaning, it is rather useless. However, sometimes a subsequence is much more frequent than the respective maximal frequent sequence, which indicates that the subsequence also appears in other contexts and, therefore, has a meaning beyond the maximal sequence context. For instance, the following two maximal frequent sequences contain clearly independent parts, which can be found if they also appear elsewhere—e.g., the name *"Oskar Lafontaine"* without the title *"finance minister"*:

finance minister Theo Waigel
finance minister Oskar Lafontaine
finance minister
Theo Waigel
Oskar Lafontaine

Preprocessing

The preprocessing phase is to transform documents into a representation that can be used by the discovery algorithm. Text includes various constituents—e.g., words, punctuation, special characters, and numbers. Basically, the discovery method accepts any sequence of these constituents and returns the frequent sequences. However, there are two reasons why some preprocessing is useful. First, some frequent words and characters may combine with other constituents in a way that can be regarded as

accidental—i.e., although the combination occurs frequently, it is more due to the frequency of the single words than due to the special meaning this combination carries. Second, the discovery method executes remarkably faster when some constituents have been removed, particularly words and characters that often occur several times within one document. Examples of these are prepositions like *"of"* or articles like *"the."* In the experiments presented in the last part of this discussion, a stoplist of 400 words was used. Furthermore, all numbers that occurred separately were removed. However, words like *"G7-countries"* remained. If the documents are in some structured form, like SGML (Standard Generalized Markup Language) or XML (Extensible Markup Language), either the structure tags have to be removed or the structure must be taken into account somehow. In the current approach, the structure is ignored.

In addition to pruning, the words can also be modified. For instance, stemming can be used to reduce inflected words to their stems. The benefit of stemming is that sequences, which have almost the same meaning but a slightly different form, are combined. If the words are not stemmed, some word sequences may be ignored, as they are not frequent enough, although the slight variations together might exceed the frequency threshold. For instance, if the following sequences both have a frequency of 7 and the frequency threshold is 10, neither of the sequences is discovered.

agricultural production
agricultural products

However, these sequences would guarantee a frequency of 14 for a stemmed sequence—e.g., *"agricultur product."*

On the other hand, stemming may join sequences that have rather different meanings. More sophisticated preprocessing may be provided by using some natural language processing like morphological analysis. Morphological analysis can reduce the words to their base forms, though at the same time retaining the knowledge on the inflected form of the word. Hence, it might be possible to combine the word forms only if the single forms are not frequent enough. This approach is not implemented yet. After all, when morphologically rich languages, like Finnish, are considered, morphological analysis clearly surpasses the stemming process. Moreover, grammatical features can be used to prune words. For instance, the discovery process may be easily restricted to nouns only. As an example of some preprocessing steps, consider the sentence

Documents are an interesting application field for data mining techniques.

The sentence can be first preprocessed using morphological analysis into the following tagged form:

(*document*_N_PL, 1) (*be*_V_PRES_PL, 2) (*an*_DET}, 3) (*interesting*_A_POS, 4)

(*application* _N_SG, 5) (*field* _N_SG, 6) (*for*_PP, 7) (*data*_N_SG, 8)
(*mining*_N_SG , 9)
(*technique*_N_PL, 10)

In the following step, uninteresting parts are pruned, resulting in

(*document* _N_PL, 1) (*interesting*_A_POS, 4) (*application* _N_SG, 5)
(*field*_N_SG, 6)
(*data*_N_SG, 8) (*mining*_N_SG, 9) (*technique*_N_PL, 10)

The text may also be pruned to contain, for example, only nouns,

(*document*_N_PL, 1) (*application*_N_SG, 5) (*field*_N_SG, 6) (*data*_N_SG, 8)
(*mining*_N_SG, 9) (*technique*_N_PL, 10)

or, depending on the goal of the research, the word itself may be discarded and only the morphological information retained:

(N_PL, 1) (A_POS, 4) (N_SG, 5) (N_SG, 6)
(N_SG, 8) (N_SG, 9) (N_PL, 10)

At the end of the preprocessing phase, the text is transformed into a long sequence with a running index. The word "###*document*###" is used to separate the documents in the sequence. The running index guarantees that each occurrence of a word has a unique location number.

Finding Maximal Frequent Word Sequences

As defined in an earlier section, maximal frequent sequences are sequences of words that appear often in the document collection and are not included in another longer frequent sequence. It is not possible to decide locally whether a sequence is frequent or not—i.e., all the instances of the sequence in the collection must be counted. Given a document with a length of twenty words, there are over 1 million possible sequences to be considered. Even if the distance of consecutive words is restricted to two, the number is still as large as 6,500. This number of candidate sequences is computationally prohibitive. Hence, a straightforward way, which includes generating the sequences for each document and then computing the sums of occurrences, is not possible in practice. Additionally, only a small fraction of sequences finally turn out to be frequent in the collection.

In the data mining community, several discovery methods have been presented. The related research includes discovery of frequent sets (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996) and discovery of sequential patterns (Agrawal & Srikant, 1995; Mannila, Toivonen, & Verkamo, 1995). In our context of textual data, a frequent set would be a set of words that co-occur frequently in documents—i.e., the order of the

words is not significant and one word may occur only once in a set. The approaches to discover sequential patterns are usually modifications of the methods for finding frequent sets. Most of these approaches use bottom-up processing: first, the frequent sets, or sequences, of size 1 are found, then longer frequent sequences are iteratively formed from the shorter ones, using the pruning principle that if a sequence is not frequent, none of the sequences in which it is included can be frequent. Finally, all the maximal sets or sequences are also found. The problem is that, when the maximal sequences are long, the number of shorter sequences is rather prohibitive. Hence, it is not possible to generate, or even consider, all the subsequences of the maximal sequences. Some approaches exist that attempt to compute the maximal sets directly; for example, in Gunopulos, Khardon, Mannila, and Toivonen (1997), a randomized algorithm is used. In Bayardo (1998), the particular goal is to handle large maximal sets. To summarize, the methods to discover sequential patterns also find maximal frequent sequences but, due to performance reasons, they do not succeed with text documents, whereas no other methods for directly finding frequent maximal sequences exist.

In the method presented in this article, the bottom-up approach is combined with the direct discovery of maximal sequences. The method is divided into two phases: initialization and discovery. First, in the initialization phase, all the frequent pairs are collected from the documents. This is done by collecting all the pairs in which the words of the pair have at most two words in between. The quantity of documents for each pair is then computed, and the pairs that occur frequently enough form the initial set of sequences for the discovery phase.

Assume the document collection contains the following six documents in which the words are replaced by letters:

1: (A,11) (B,12) (C,13) (D,14) (E,15)
2: (P,21) (B,22) (C,23) (D,24) (K,25)
3: (A,31) (B,32) (C,33) (H,34) (D,35) (K,36)
4: (P,41) (B,42) (C,43) (D,44) (E,45) (N,46)
5: (P,51) (B,52) (C,53) (K,54) (E,55) (L,56) (M,57)
6: (R,61) (H,62) (K,63) (L,64) (M,65)

In order to be counted as an instance of a pair in the initialization phase, two words in a document can have between them at most two other words. Assume that the frequency threshold is set to 2:

AB: 2 BE: 3 CK: 3 EL: 1 HM: 1 PC: 3
AC: 2 BH: 1 CL: 1 EM: 1 KE: 1 PD: 2
AD: 1 BK: 2 CN: 1 EN: 1 KL: 2 PK: 1
AH: 1 CD: 4 DE: 2 HD: 1 KM: 2 RH: 1
BC: 5 CE: 3 DK: 2 HK: 2 LM: 2 RK: 1
BD: 4 CH: 1 DN: 1 HL: 1 PB: 3 RL: 1

As longer sequences are constructed from the shorter ones, a container of the current sequences has to be maintained. This container is called the *set of grams*, and the sequences contained in this set are called *k*-grams in which *k* is the length of the gram. After the initialization phase, the set of grams contains all the pairs—i.e., 2-grams that have a frequency of at least 2: {AB, AC, BC, BD, BE, BK, CD, CE, CK, DE, DK, HK, KL, KM, LM, PB, PC, PD}.

The discovery phase, in turn, is divided into three steps: the expansion, the join, and the prune steps. In the expansion step, a new word is added to the pair and, if the resulting 3-gram is still frequent, the same is repeated until the resulting sequence is no more frequent. The last frequent sequence is now a maximal frequent sequence.

Continuing with the previous example, the processing starts from the pair AB. The set of grams is used to find candidate words to be added since only sequences that are constructed from the frequent pairs can be frequent. A candidate sequence ABC can be constructed using the grams, and its frequency, when checked against the documents, is found to be 2. A new attempt is made to add a word: the sequence ABCD is again found to be frequent. However, all other attempts to make the sequence longer fail since the possible candidate sequences ABCDE and ABCDK are not frequent. Hence, the sequence ABCD is a maximal frequent sequence.

The other pairs are processed in a similar way. However, to ensure that the same maximal sequences are not produced several times, a pair is not expanded if it is already a subsequence of some maximal sequence. When all the pairs have been processed, every pair belongs to some maximal sequence. If some pair cannot be expanded, it is itself a maximal sequence. A pair that is a maximal sequence can be removed from the set of grams as it cannot be a part of any other maximal sequence.

In the previous example, the pairs AC, BC, and BD are subsequences of ABCD and, hence, are not expanded. The pair BE can be expanded to the sequence BCDE, BK to BCDK, KL to KLM, and PD to PBCD. The pair HK cannot be expanded, and it is removed from the set of grams. All other pairs are subsequences of some maximal sequence. Hence, after the first expansion step, the set of maximal frequent sequences is {ABCD, BCDE, BCDK, KLM, PBCD, HK}.

In the expansion step, all the possibilities to expand may have to be checked—i.e., the new word can be added to the end, to the front, or to the middle of the sequence. If one expansion does not produce a frequent sequence, other alternatives have to be considered. However, when a suitable word is found, other alternatives can be ignored.

As seen earlier, the test for frequency of some sequence has to be done regularly. Due to efficiency reasons, it is not reasonable to check the frequency from the documents. Therefore, the occurrences of the frequent pairs found from the documents are stored in a separate data

structure. For instance, for the pairs AB, AC, and BC, the following occurrences are stored:

AB: [11-12][31-32]
AC: [11-13][31-33]
BC: [12-13][22-23][32-33][42-43][52-53]

If the frequency of the sequence ABC has to be found, the data structure is first searched to find the occurrences of AB, in this case [11-12] and [31-32]. Then those occurrences of BC that are continuations of the occurrences of AB are found, namely [12-13] and [32-33]. Finally the occurrences for the entire sequence ABC can be combined: [11-13] and [31-33]. The number of these occurrences is returned as the frequency of the sequence.

In the join step, the pairs—i.e., 2-grams—are joined to form 3-grams. For instance, if there are pairs AB, BC, and BD, new sequences ABC and ABD are formed. Within the example, the following 3-grams are included in the new set of grams, but the 3-gram KLM can be removed since it is already known to be maximal:

ABC	ACK	CDE	PCD	BKM
ABD	BCD	CDK	PCE	CKL
ABE	BCE	PBC	PCK	CKM
ABK	BCK	PBD	PDE	DKL
ACD	BDE	PBE	PDK	DKM
ACE	BDK	PBK	BKL	(KLM)

Now the expansion step is repeated: all 3-grams are processed and expanded. The difference, however, is that the 3-grams are not necessarily frequent. Actually, a 3-gram formed by joining two 2-grams does not necessarily occur at all in the documents. For instance, if there are pairs AB and BC but BC always occurs before AB in the documents, the sequence ABC never occurs. The 3-grams that are not frequent are pruned away from the set of grams.

The 3-grams ABC and ABD are subsequences of ABCD, which is maximal. The 3-gram ABE occurs only once in the documents and can be removed from the set of grams. In fact, there are only two 3-grams, namely PBE and PBK, that are frequent but are not already subsequences of maximal sequences. These grams are expanded to maximal sequences PBCE and PBCK, respectively.

After the expansion step, the set of grams looks like the following:

ABC	BCE	CDE	PBE	PCK
ABD	BCK	CDK	PBK	
ACD	BDE	PBC	PCD	
BCD	BDK	PBD	PCE	

The expansion and join steps are iterated in a similar way as explained above: *k*-grams are expanded to maximal sequences and then joined to

form *k+1*-grams. As all the grams that are itself maximal sequences are removed, as well as the grams that are not frequent, the processing stops when the set of grams is empty. In practice, when the sequences can be long (e.g., twenty-two words), it is necessary to apply pruning, which makes it possible to remove grams before the gram length is the same as the length of the maximal sequence. The prune step is explained shortly below.

The following 4-grams belong to the set of grams. The grams ABCD, BCDE, BCDK, PBCD, PBCE, and PBCK are maximal, the other grams become infrequent. Hence, the processing stops after this expansion step:

(ABCD) ABDK (BCDK) PBDE
ABCE ACDE (PBCD) PBDK
ABCK ACDK (PBCE) PCDE
ABDE (BCDE) (PBCK) PCDK

Finally, the set of maximal frequent sequences includes the sequences {*ABCD, BCDE, BCDK, KLM, PBCD, HK, PBCE, PBCK*}. All of these maximal sequences have a frequency of 2.

The basic version of the algorithm proceeds to as many levels as is the length of the longest maximal sequence. When there exist long maximal sequences, this can be prohibitive, since on every level the join operation increases exponentially the number of the grams contained in the maximal sequences. However, often it is unnecessary to wait until the length of the grams is the same as the length of the maximal sequence in order to remove the grams of the maximal sequence from the set of grams. As there must always be grams from at least two maximal sequences to form a new maximal sequence, we can check, for each maximal sequence *m*, whether there are any other maximal sequences that can still contribute in forming new sequences and, hence, can preclude the removal of the grams of *m*. If no such other maximal sequences are found, the maximal sequence is declared to be *final*.

For instance, the following principles can be used on the level *k* to decide whether a maximal sequence is final:

- A maximal sequence is final if it does not share any *k-1*-prefixes or suffixes of grams with another maximal sequence in at least *n* documents in which *n* is the frequency threshold.
- Let *M* be the set of maximal sequences that share a *k-1*-prefix or -suffix of a gram with the maximal sequence *m*. For each *m'* in *M*, we align *m* and *m'* to know which subsequences of *m* are also subsequences of *m'*. If there is a common subsequence which is longer than the maximal gap, the sequence *m' binds m*. Otherwise, if there are no subsequences, such that the respective positions in *m* have a distance larger than *k*, the sequence *m'* does not *bind m*. If no *m' in M* binds *m*, *m* is final.

The pruning of maximal sequences proceeds as follows. After a join step on level k, as the set of grams contains $k+1$-grams, the maximal sequences are processed. First, if a maximal sequence is of size $k+1$, the respective gram is removed from the set of grams. Second, for all the longer maximal sequences, it has to be decided whether the sequence is final. Finally, after all the maximal sequences have been considered, any gram that is a subsequence of final maximal sequences only is removed. Only the grams of a maximal sequence are removed; the sequence is not removed from the set of maximal sequences in order to assure that the new sequences are not already subsequences of some maximal sequence.

Above, a method for discovering maximal frequent sequences was presented. After all the maximal frequent sequences have been discovered, their subsequences that are more frequent than the corresponding maximal sequence can also be found. The more frequent subsequences are found for each maximal sequence in turn. Continuing the running example, first, a set of grams is formed from the pairs contained in the maximal sequence ABCD—i.e., the set contains the 2-grams {AB, AC, AD, BC, BD, CD}. Second, two grams are joined to form a 3-gram if the frequency of the 3-gram is higher than the frequency of the maximal sequence—i.e., in this case, higher than 2. Only one such 3-gram can be formed—i.e., BCD, which has a frequency of 4. As a subsequence can also be a subsequence of some other subsequence of the maximal sequence, it is only considered interesting if it is more frequent than any of the other subsequences that contain it. Hence, a 2-gram is included in the set of more frequent subsequences if it is more frequent than the maximal sequence, and either it is not included in any 3-gram or it is more frequent than any of the 3-grams in which it is contained. From the 2-grams in the set of grams above, only the gram BC fulfills the requirements. The grams BD and CD are more frequent than the maximal sequence ABCD, but these are contained in BCD and have the same frequency as it has, namely 4. These steps, joining of the k-grams to form k+1-grams and checking whether the k-grams fulfill the requirements, are repeated until the set of grams is empty. This procedure is done for each maximal frequent sequence. Within the previous example, the final set of more frequent subsequences looks like the following (the number in parentheses is the frequency of the sequence):

ABCD(2): BCD(4), BC(5)
BCDE(2): BCD(4), BCE(3), BC(5)
BCDK(2): BCD(4), BCK(3), BC(5)
KLM(2): no more frequent subsequences
PBCD(2): PBC(3), BCD(4), BC(5)
HK(2): no subsequences
PBCE(2): PBC(3), BCE(3), BC(5)
PBCK(2): PBC(3), BCE(3), BC(5)

Examples of both maximal frequent sequences and their more frequent subsequences, as extracted from a newswire text collection, can be found in the Examples and Experiments section.

Ordering and Pruning Word Sequences

Some decisions will now be presented which reduce the number of word sequences found in the documents. First, to be counted as a word sequence, the sequence cannot contain gaps that are longer than the maximal gap given. Second, the word sequence has to be frequent. Third, a word sequence has to be a maximal sequence or a subsequence of some maximal sequence such that the subsequence has to be more frequent than the respective maximal sequence. However, even after these restrictions, the number of word sequences found may be rather large, at least for some application purposes. In this section, some postprocessing opportunities are presented that make it possible to find an ordering for the word sequences. If the word sequences are ordered according to their assumed quality, the less valuable sequences can be pruned if necessary. Also, even a large set of sequences is easier to utilize if the sequences are presented in the order of their assumed quality.

A central problem with the knowledge discovery approaches is to define the notion of *interestingness.* The basic methods usually return large sets of information, and some extra effort is needed to evaluate the results and decide which are interesting for the given application.

With the approach presented in the previous sections, interestingness is in the first place determined by the frequency of the sequence. Actually, with long sequences, it is a very good measure since the frequency guarantees that the combination is not accidental. On the other hand, long sequences hardly appear so often that they lose their interestingness value. However, with single words and short sequences, it may happen that a high frequency indicates that the sequence has no describing power and that its interestingness is low. Hence, some further measures have to be considered to focus on good sequences.

In data mining, typically the relevance of a pattern is very strongly dependent on the application. One of our hypotheses is that there are measures of relevance that are common to all documents independent of the semantic content of the texts. However, different usage needs also affect application of the measures. In the following, a set of measures is introduced. Each of these measures attempts to describe one aspect of interestingness using knowledge gathered from the word sequences. The measures are:

- **Length:** The absolute length *wlen* of a word sequence is the number of words it contains. The relative length *len* is defined as *1 - (1/wlen).*

- **Frequency:** The frequency *f* of a word sequence is the number of documents in which the sequence occurs. We also consider separately the number of occurrences of a word sequence within a document. This frequency is denoted as *tf*.
- **Tightness:** If the first word of a sequence appears in location *x* and the last word in location *y*, the window of the occurrence is defined as $y - x + 1$. If *win* is the average of all the windows of the occurrences of a sequence, the tightness *t* of the sequence is $1 - ((win - wlen)/win)$.
- **Stability**: A word sequence *s* of absolute length *n* contains $P = \sum_{k=2}^{n-1} \binom{n}{k}$ subsequences.
  Assume there are subsequences $s_1, \ldots, s_k$ with the frequencies higher than the frequency of *s*, and for each s_i, $cont(s_i)$ is the number of subsequences s_i itself contains. Additionally, let *S* be the sum of these sets, i.e., $S = cont(s_1) + \ldots + cont(s_k)$. Furthermore, the relative frequency of these subsets is computed as $F = cont(s_1) \times f(s)/f(s_1) + \ldots + cont(s_k) \times f(s)/f(s_k)$ in which $f(s_i)$, $1 \leq i \leq k$, are the frequencies of the subsequences $s_1, \ldots, s_k$, respectively. The stability *st* of the word sequence *s* is now computed as $st = (F/S + (1 - S/P))/2$.
- **Inverse Document Frequency (IDF):** The IDF is computed as $idf = -\log$ (number of documents containing the sequence / number of documents in the collection). This measure is also scaled to receive values between $0 - 1$. The relative IDF is computed by dividing *idf* by the maximal IDF value *max_idf*, which is obtained when a word sequence occurs in *n* documents, in which *n* is the frequency threshold—i.e., $rel\_idf = idf / max\_idf$.

In order to illustrate the measures of tightness and stability, which are not as intuitive as the other measures, some examples are given in the following:

Tightness depicts the number of gaps that can be found, on the average, in the occurrences of the word sequence in the documents. All the occurrences are counted—i.e., also the duplicate occurrences in one document. In the first example of this article, three documents were presented. In these documents, a word sequence *(unfair, practices)* occurred in locations (83, 86), (118, 120), and (420, 421). The average window is now 3, and the length of the sequence is 2. Hence, the tightness is $1 - (3-2)/3 = 1 - 1/3 = 0.67$.

Stability attempts to reveal the ties between the constituents of the word sequence. If the subsequences of the sequence occur in the context of the sequence only, which is depicted by the fact that they have the same frequency as the sequence itself, the stability is 1. If, however, there are subsequences that are more frequent, the stability gets a value between 0 and 1.

In the following, one word sequence and its more frequent subsequences with frequencies are presented:

available export enhancement program initiative announced 11
—program initiative 22
—program announced 37
—enhancement initiative 25
—enhancement announced 20
—available export 25
—export initiative announced 19

As the length of the sequence is 6, the sequence has 56 subsequences—i.e., $P = 56$. Only subsequences longer than 2 are counted. Therefore, from the more frequent subsequences, only the last one, which itself has a length of 3, has subsequences of its own. The sum of more frequent subsequences is 8—i.e., $S = 8$. $F = 11/22 + 11/37 + 11/25 + 11/20 + 11/25 + 3 \times 11/19 = 3.97$. The stability is now computed as $st = (3.97/8 + (1 - 8/56)) / 2 = 0.68$.

The ranking of word sequences within a document is computed by combining the above measures. It can be done in several ways, depending on the emphasis given to each measure. As the frequency is considered to be the central factor, the weight w is computed by multiplying the frequency of a word sequence within a document by a combination of the other measures. In the experiments presented in this article, the weight is computed as $w = tf \times (len + t + st + rel\_idf) / 4$. After each word sequence has a weight, it is possible to select the best word sequences and prune away the sequences with a weight less than a given threshold. The threshold can have a predefined fixed value or a dynamic value depending on, for example, the number of the word sequences found in the collection.

Use of Frequent Word Sequences

In this section, some possible uses of frequent word sequences are discussed. These uses include applying word sequences within some existing applications, but this technology may also make available new ways to face information overload.

For instance, imagine the following situation. The user makes a query that may contain single words or a more complicated search expression. Instead of returning the documents that fulfill the conditions of the query, the search engine returns only the maximal frequent word sequences. The user can see the contexts of the query terms and use this knowledge to decide which maximal sequences correspond to his or her information need. Assume the user has given a query term *"agriculture."* If stemming is used, the following word sequences might be returned:

agricultural exports
agricultural production

agricultural products
agricultural stabilization conservation service
agricultural subsidies
agricultural trade
u.s. agriculture
agriculture department usda
agriculture department wheat
agriculture hearing
agriculture minister
agriculture officials
agriculture subcommittee
agriculture undersecretary daniel amstutz
common agricultural policy
ec agriculture ministers
european community agriculture
food agriculture
house agriculture committee

If the user chooses the sequence *"agricultural subsidies,"* the search engine can return the documents containing this sequence. Instead of choosing one sequence only, also several sequences, connected by Boolean operators, can be chosen.

Before seeing the entire documents, the user could also ask, Which other word sequences appear in the same documents with the sequence *"agricultural subsidies?"* That is, instead of the text of the documents, the set of word sequences for each document would be returned. This representation would give the user an overview to the documents, particularly if not only the plain word sequence is returned but also the entire sentence or paragraph in which the word sequence occurs.

Furthermore, the user could choose some documents to be viewed. If some document looks promising, the entire document can be used as a query to search other documents that are somehow similar. The similarity comparison that is needed to implement this functionality can be based on the frequent word sequences as a representation of a document.

In particular, the word sequences might support information seeking when the user does not have a very clear idea what he or she is looking for, and when the user is not familiar with the document collection. The frequent word sequences offer a mediating level, which can be used to expand even very short queries. As today ordinary users can directly search large full-text document collections without the help of a professional librarian, new challenges are to be addressed. According to some studies (Grefenstette, 1997), users tend to use few words only in their queries even if search engines offer more advanced search expressions. Hence, new simple tools should be offered users. The word sequences are close to the normal text and probably easy to accept.

In addition to the information searching process where the user has even some kind of an idea what he or she is looking for, the word sequences also support knowledge discovery where the user is looking for

new, interesting, and surprising knowledge contained in the documents. For instance, a discovery process might find associations between word sequences—e.g., that names of two companies seem to occur more often in the same documents than what is to be expected. Moreover, the documents can be clustered using the word sequences. Clusters give an overview to the collection, and these can also be a way to restrict the scope of further searching to a subset of the documents. Finally, as the word sequences inherently have exactly the same form in several documents, they can be used to automatically generate hypertext links between documents.

In practice, the set of word sequences for a document should probably be expanded by frequent content-bearing single words of the document. Some words that have an important meaning occur in many contexts, which may preclude them participating in frequent word sequences. Furthermore, in many languages other than English, composite words are favored instead of phrases of several words.

Examples and Experiments

The publicly available Reuters-21578 newswire collection, which contains about 19,000 documents, has been used for experiments (see http://www.research.att.com/~lewis/reuters21578.html). The average length of the documents is 135 words. Originally, the documents contained 2.5 million words. After stopword pruning against a list of 400 stopwords, the amount of words is reduced to 1.3 million, which are instances of 50,000 distinct words.

The list of stopwords contains, for example, single letters, pronouns, prepositions, some common abbreviations from the Reuters collection (e.g., pct, dlr, cts, shr), and some other common words that do not have any content-bearing role in text phrases. On the other hand, many common verbs are not stopwords since they may add some necessary context into phrases (e.g., call, bring, and sell). In addition to stopwords, all numbers are removed from the documents. Stopword pruning is necessary, since the efficiency of the algorithm may be ruined if there are many words that appear several times in one document. Frequency of words as such is not a problem, although they naturally increase the size of data, and sequences containing very frequent words only may not be very interesting.

Experiments have been performed with frequency thresholds 10 and 15. Table 1 summarizes how many pairs and occurrences were found in the initial phase.

Table 1.
Frequent Pairs and Number of Their Occurrences

Frequency	10	15
Frequent pairs	22,273	2,071
Occurrences	556,522	427,775
Pairs/document (average)	30	23

Additionally, there were 406 documents without any frequent pairs. Table 2 shows some performance figures of the execution with a frequency threshold of 10. Before the first pass over the set of grams, 3.40 minutes were needed to construct the initial data structures. As can be seen from the table, most of the time is used on the first level (actually level 2), when most of the maximal sequences, also the longer ones, are extracted. Although the longest maximal sequences are twenty-five words long, the process ends after the fourth round. Hence, the pruning of grams of the maximal sequences is effective.

Table 2.
TIME AND SPACE CONSUMPTION WITH A FREQUENCY 10

Level	2	3	4
Pass time over the set of grams (min)	44.30	2.33	0.06
Join time (min)	7.27	0.19	0.00
Grams in the set	22,273	4,849	152
Grams that not-in-maximal when seen	17,089	3,898	136
New maximal sequences	17,089	796	32
Grams that are maximal	13,984	709	31
Nonfrequent grams	0	3,102	104
Final maximal sequences	17,007	877	33
Maximal sequences remain	82	1	0
Grams after *join*	12,042	1,544	21
Grams after pruning	4,849	152	0

With a frequency 15, the total time of the execution was 30.42 minutes and the algorithm proceeded, as with the frequency 10, through levels 2, 3, and 4. The numbers of frequent maximal sequences of various sizes are shown in Table 3.

Table 3.
NUMBERS OF MAXIMAL FREQUENT SEQUENCES OF DIFFERENT LENGTH

Length	2	3	4	5	6	7	8	9	10	11	12	13
σ=10	43,273	8,417	1,963	764	321	116	38	34	21	18	15	10
σ=15	21,635	4,169	1,002	394	171	62	22	19	10	9	7	6

Length	14	15	16	17	18	19	20	21	22	23	24	25
σ=10	4	2	2		16	2	2	1	4			2
σ=15	2		1	1		8	1	1	1	2		

As an example of word sequences discovered for one document, consider the following sequences:

power station	11
immediately after	26
co operations	11
effective april	63
company's operations	20
unit nuclear	12
unit power	16
early week	42
senior management	28
nuclear regulatory commission	14
—regulatory commission	34
nuclear power plant	26
—power plant	55
—nuclear power	42
—nuclear plant	42
electric co	143

The document describes how the Nuclear Regulatory Commission ordered one nuclear power plant to be shut down after determining that operators were sleeping on duty. Moreover, the action of the electric company to improve the situation is stated. As the discovery of word sequences is based on frequency, the word sequences extracted tell more about the domain than about the details covered in this specific story.

In addition to the discovery of maximal frequent word sequences, more frequent subsequences were also found. Instead of accepting subsequences that are only slightly more frequent, it was required that a subsequence occur at least in seven documents more than for the corresponding maximal sequence. In the experiments, the number of maximal sequences that had subsequences that were more frequent than the maximal sequence itself was 2,264 within the average of two subsequences. As an example of a long word sequence that has several more frequent subsequences, consider the following sequence:

> previous business day treasury latest budget statement
> balances tax loan note accounts fell respective days
> treasury's operating cash balance totaled

The frequency of the sequence is 13 and contains the following subsequences that are more frequent:

operating cash	29
cash balance	29
business day	33
treasury statement	24
statement tax	22
statement loan	25
latest statement	23
previous day	25
previous treasury	21
budget statement	22

budget tax	37
tax loan	23
tax note	27

Finally, for each word sequence, the measures described in an earlier section were computed. In Table 4, the measures for the document already discussed can be seen. The word sequence *nuclear regulatory commission*, and hence also the sequence *regulatory commission*, occurs in the document twice, which contributes to the high weight. It can be seen from the tightness (*t*) values that the sequences *power station*, *nuclear regulatory commission*, and *nuclear power* are very rigid compositions, as they never allow other words within them. The sequence *nuclear power plant* receives rather poor stability (*st*) value, since all of its subsequences are clearly more frequent. Furthermore, the sequence *electric co* has the worst relative IDF (*rel_idf*) value due to its frequency in the document collection.

Table 4.
WEIGHTING MEASURES FOR WORD SEQUENCES OF ONE DOCUMENT (F = FREQUENCY, T = TIGHTNESS, ST = STABILITY, REL_IDF = RELATIVE INVERTED DOCUMENT FREQUENCY, LEN = LENGTH)

Word Sequence	*f*	*t*	*st*	*rel_idf*	*len*	*weight*
power station	11	1.00	-	0.99	0.50	0.83
immediately after	26	0.98	-	0.87	0.50	0.78
co operations	11	0.63	-	0.99	0.50	0.71
effective april	63	0.98	-	0.75	0.50	0.74
company's operations	20	0.65	-	0.91	0.50	0.68
unit nuclear	12	0.67	-	0.98	0.50	0.71
unit power	16	0.55	-	0.94	0.50	0.66
early week	42	0.92	-	0.81	0.50	0.74
senior management	28	0.98	-	0.86	0.50	0.78
nuclear regulatory commission	14	1.00	0.54	0.95	0.67	1.58
- regulatory commission	34	0.99	-	0.84	0.50	1.54
nuclear power plant	26	0.99	0.29	0.87	0.67	0.70
- power plant	55	0.93	-	0.77	0.50	0.74
- nuclear power	42	1.00	-	0.81	0.50	0.77
- nuclear plant	42	0.77	-	0.81	0.50	0.69
electric co	143	0.86	-	0.65	0.50	0.67

CONCLUSION

In this article, a method for extracting maximal frequent word sequences from documents was presented, and its possible uses in several tasks were discussed. A maximal frequent word sequence is a sequence of words that is frequent in the document collection and, moreover, is not contained in any longer frequent sequence. A sequence is considered to be frequent if it appears in at least n documents, when n is the frequency threshold given.

The process for extracting useful word sequences contains several phases. In the preprocessing phase, some common words and numbers are pruned from documents, and the documents are transformed into a sequence of words. In the discovery phase, on the one hand, longer word sequences are constructed from shorter ones; on the other hand, maximal sequences are discovered directly in order to avoid a prohibitive number of intermediate sequences. Subsequences which are more frequent than the corresponding maximal sequence can also be found in a separate postprocessing step. When each document has received a set of frequent word sequences, the sequences are ordered according to a set of measures that assess the quality of the sequences. If necessary, the set of sequences can be pruned based on the ordering.

The representation of a document as a set of frequent word sequences can form a basis for several different information searching tools. First, the word sequences can serve as content descriptors that can be used for similarity computations needed—e.g., in clustering, automatic hypertext link generation, and matching documents with a query. Second, after a user query, the word sequences can be shown as an intermediate representation of the documents before the user makes the final decision to see the entire texts. Third, the word sequences may act as a set of features for further knowledge discovery. For instance, associations between word sequences, and hence between documents, can be discovered.

The discovery method and ordering principles have been implemented in the Perl programming language. In experiments, the Reuters-21578 newswire collection was used. Further research in the area includes applying the method to the information searching tasks described earlier.

REFERENCES

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; & Verkamo, A. I. (1996). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 307-328). Menlo Park, CA: AAAI Press.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering* (March 6-10, 1995, Taipei, Taiwan) (pp. 3-14). Los Alamitos, CA: IEEE Computer Society Press.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. In L. Haas & A. Tiwary (Eds.), *SIGMOD '98* (Proceedings of the 1998 ACM SIGMOD Conference on Management of Data: June 1-4, 1998, Seattle, Washington) (pp. 85-93). New York: Association for Computing Machinery Press.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (August 1995, Montreal, Canada) (pp. 112-117). Menlo Park, CA: AAAI Press.

Feldman, R.; Dagan, I.; & Klösgen, W. (1996). Efficient algorithms for mining and manipulating associations in texts. In R. Trappl (Ed.), *Cybernetics and systems research: The Thirteenth European Meeting on Cybernetics and Systems Research* (pp. 949-954). Vienna, Austria: Taylor & Francis.

Grefenstette, G. (1997). Short query linguistic expansion techniques: Palliating one-word queries by providing intermediate structures to text. In M. T. Pazienza (Ed.), *Information extraction: A multidisciplinary approach to an emerging information technology* (No. 1299: Lecture Notes in Artificial Intelligence) (pp. 97-114). New York: Springer.

Gunopulos, D.; Khardon, R.; Mannila, H.; & Toivonen, H. (1997). Data mining, hypergraph transversals, and machine learning. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-*SIGART Symposium on Principles of Database Systems, PODS 1997 (Tuscon, Arizona, May 12-14, 1997) (pp. 209-216). New York: Association for Computing Machinery Press.

Hearst, M. A. (1995). Tilebars: Visualization of term distribution information in full text information access. In *CHI '95: Human Factors in Computing Systems: "Mosaic of Creativity"* (May 7-11, 1995, Denver, Colorado) (pp. 59-66). New York: Association for Computing Machinery.

Heinonen, O. (1998). Optimal multi-paragraph text segmentation by dynamic programming. In *COLING-ACL '98* (36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, COLING-ACL '98) (pp. 1484-1486). Montreal, Canada: Université de Montreal.

Mannila, H.; Toivonen, H.; & Verkamo, A. I. (1995). Discovering frequent episodes in sequences. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (August 1995, Montreal, Canada) (pp. 210-215). Menlo Park, CA: AAAI Press.

Template Mining for Information Extraction from Digital Documents

GOBINDA G. CHOWDHURY

ABSTRACT

WITH THE RAPID GROWTH OF DIGITAL INFORMATION RESOURCES, information extraction (IE)—the process of automatically extracting information from natural language texts—is becoming more important. A number of IE systems, particularly in the areas of news/fact retrieval and in domain-specific areas, such as in chemical and patent information retrieval, have been developed in the recent past using the template mining approach that involves a natural language processing (NLP) technique to extract data directly from text if either the data and/or text surrounding the data form recognizable patterns. When text matches a template, the system extracts data according to the instructions associated with that template. This article briefly reviews template mining research. It also shows how templates are used in Web search engines—such as Alta Vista—and in meta-search engines—such as Ask Jeeves—for helping end-users generate natural language search expressions. Some potential areas of application of template mining for extraction of different kinds of information from digital documents are highlighted, and how such applications are used are indicated. It is suggested that, in order to facilitate template mining, standardization in the presentation and layout of information within digital documents has to be ensured, and this can be done by generating various templates that authors can easily download and use while preparing digital documents.

Gobinda G. Chowdhury, Division of Information Studies, School of Applied Science, Nanyang Technological University, N4#2a—32, Nanyang Avenue, Singapore 639798
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 182-208

Information Extraction and Template Mining

Information extraction (IE), the process of automatically extracting information from natural language texts, is gaining more and more importance due to the fast growth of digital information resources. Most work on IE has emerged from research into rule-based systems in natural language processing. Croft (1995) suggested that IE techniques, primarily developed in the context of the Advanced Research Projects Agency (ARPA) Message Understanding Conferences (MUCs), are designed to identify database entities, attributes, and relationships in full text. Gaizauskas and Wilks (1998) defined IE as the activity of automatically extracting pre-specified sorts of information from short natural language texts typically, but by no means exclusively, newswire articles. Although works related to IE date back to the 1960s, perhaps the first detailed review of IE as an area of research interest in its own right was by Cowie and Lehnert (1996). However, a detailed review dividing the literature on IE into three different groups—namely, the early work on template filling, the Message Understanding Conferences (MUCs), and other works on information extraction—has recently been published by Gaizauskas and Wilks (1998).

Template mining is a particular technique used in IE. Lawson et al. (1996) defined template mining as a natural language processing (NLP) technique used to extract data directly from text if either the data and/or text surrounding the data form recognizable patterns. When text matches a template, the system extracts data according to instructions associated with that template. Although different techniques are used for information extraction and knowledge discovery—as described by Cowie and Lehnert (1996), Gaizauskas and Wilks (1998), and Vickery (1997)—template mining is probably the oldest information extraction technique. Gaizauskas and Wilks (1998) reported that templates were used to extract data from natural language texts against which "fact retrieval" could be carried out in the Linguistic String Project at New York University that began in the mid-1960s and continued into the 1980s (reported by Sager, 1981). Numerous studies have been conducted, though most of them are domain-specific, using templates for extracting information from texts. This article briefly reviews some of these works. It also shows how templates are used for information retrieval purposes in major Web search engines like AltaVista (http://www.altavista.com). This discussion proposes that template mining has great potential in extracting different kinds of information from documents in a digital library environment. To justify this proposition, this article reports some preliminary tests carried out on digital documents, more specifically on some articles published in the *D-Lib Magazine* (http://www.dilib.org/dilib).

Works on Template Mining

Template mining has been used successfully in different area:

- extraction of proper names by Coates-Stephens (1992), Wakao et al. (1996), and by Cowey and Lehnert (1996);
- extraction of facts from press releases related to company and financial information in systems like ATRANS (Lytinen & Gershman, 1986), SCISOR (Jacobs & Rau, 1990), JASPER (Andersen, et al., 1992; Andersen & Huettner, 1994), LOLITA (Costantino, Morgan, & Collingham, 1996), and FIES (Chong & Goh, 1997);
- abstracting scientific papers by Jones and Paice (1992);
- summarizing new product information by Shuldberg et al. (1993);
- extraction of data from analytical chemistry papers by Postma et al. (1990a, 1990b) and Postma and Kateman (1993);
- extraction of reaction information from experimental sections of papers in chemistry journals by Zamora and Blower (1984a,1984b);
- processing of generic and specific chemical designations from chemical patents by Chowdhury and Lynch (1992a,1992b) and by Kemp (1995); and
- extraction of bibliographic citations from the full texts of patents by Lawson et al. (1996).

Template mining has largely been used for extraction of information from news sources and from texts in a specific domain. Gaizauskas and Wilks (1998) reported that applied work on filling structured records from natural language texts originated in two long-term research projects: The Linguistic String project (Sager, 1981) at New York University and the research on language understanding and story comprehension carried out at Yale University by Schank and his associates (Schank, 1975; Schank & Abelson, 1977; Schank & Riesbeck, 1981). The first research was conducted in the medical science domain, particularly involving radiology reports and hospital discharge summaries, while the second research led to many other research works in the early 1980s that used the principles and techniques of IE to develop practical applications such as the FRUMP system developed by De Jong (1982). FRUMP used a simplified version of SCRIPTS, proposed by Schank (Schank, 1975; Schank & Abelson, 1977; Schank & Riesbeck, 1981), to process text from a newswire source to generate story summaries.

ATRANS (Lytinen & Gershman, 1986), another IE system, was soon developed and commercially applied. ATRANS used the *script* approach (Schank & Abelson, 1977; Schank & Riesbeck, 1981) for automatic processing of money transfer messages between banks. Another successful application of IE has produced a commercial online news extraction system called SCISOR (Jacobs & Rau, 1990) that extracts information about corporate mergers and acquisitions from online news sources. JASPER

(Andersen et al., 1992; Andersen & Huettner, 1994) was another IE system developed for fact extraction for Reuters. JASPER uses a template-driven approach and partial analysis techniques to extract certain key items of information from a limited range of texts such as company press releases. LOLITA (Costantino, Morgan, & Collingham, 1996) is a financial IE system that uses three pre-defined groups of templates designed according to a "financial activities approach," namely, company related templates, company restructuring templates, and general macroeconomic templates. In addition, the user-definable template allows the user to define new templates using natural language sentences. Chong and Goh (1997) developed a similar template-based financial information extraction system, called FIES, that extracts key facts from online news articles.

Applications of template mining techniques for automatic abstracting can be traced back to 1981 when Paice (1981) used what he called *indicator phrases* (such as "the results of this study imply that . . . ") to extract topics and results reported in scientific papers for generating automatic abstracts. Paice continued his work to improve on this technique and for resolving a number of issues in natural language processing (see, for example, Jones & Paice, 1992; Paice & Husk, 1987). Shuldberg et al. (1993) described a system that digests large volumes of text, filtering out irrelevant articles and distilling the remainder into templates that represent information from the articles in simple slot/filler pairs. The system consists of a series of programs each of which contributes information to the text to help determine which strings constitute appropriate values for the slots in the template.

Chemical and patent information systems have been the prominent areas for the application of templates for IE. TICA (Postma et al., 1990a, 1990b; Postma & Kateman, 1993) used templates to extract information from the abstracts of papers on inorganic tritimetric analysis. The parsing program used in TICA followed an expectation-driven approach where words or groups of words expect other words or concepts to appear. Zamora and Blower (1984a,1984b) developed a system that automatically generates reaction information forms (RIFs) from the descriptions of syntheses of organic chemicals in the *Journal of the American Chemical Society*. The techniques explored in the semantic phase of this work include the use of a *case grammar* and *frames* (Schank & Abelson, 1977; Schank & Riesbeck, 1981) to map the surface structure of the text into an internal representation from which the RIFs can be formed. Following the same methodology, Ai et al. (1990) developed a system that generates a summary of all preparative reactions from the experimental sections of the *Journal of Organic Chemistry* papers. This work identified seven sequences of events that were used for building templates for the text of an experimental paper.

Chowdhury and Lynch (1992a, 1992b) developed a template-based method for converting to GENSAL (a generic structure language developed

at the University of Sheffield) those parts of the Derwent Documentation Abstracts that specify generic chemical structures. Templates for processing both the variable and multiplier expressions, which predominate in the assignment statements in the Derwent Documentation Abstracts, were identified for further processing. As part of this research, Chowdhury (1992) also conducted a preliminary discourse analysis of European chemical patents that identified the common patterns of expressions occurring in different parts of patent texts. This work prompted further research (Kemp, 1995; Lawson et al., 1996) leading to the use of template mining in the full text of chemical patents. Lawson et al. (1996) reported their work using the template mining approach to isolate and extract automatically bibliographic citations to patents, journal articles, books, and other sources from the full texts of English-language patents.

There is also some work that examines the development of specific tools and techniques for information extraction using templates. For example, Sasaki (1998) reported an ongoing project on building an information extraction system that extracts information from a real-world text corpus such as newspaper articles and Web pages. As part of this project, an inductive logic programming (ILP) system has been developed to generate IE rules from examples. Gaizauskas and Humphreys (1997) described the approach taken to knowledge representation in the LaSIE information extraction system, particularly the knowledge representation formalisms, their use in the IE task, and how the knowledge represented in them is acquired. LaSIE first translates individual sentences to a quasi logical form and then constructs a discourse model of the entire text from which template fills are derived.

Guarino (1997) argued that the task of information extraction can be seen as a problem of semantic matching between a user-defined template and a piece of information written in natural language. He further suggested that the ontological assumptions of the template need to be suitably specified and compared with the ontological implications of the text. Baralis and Psaila (1997) argued that the current approaches to data mining usually address specific user requests, while no general design criteria for the extraction of association rules are available for the end-user. To solve this problem, they have proposed a classification of association rule types that provides a general framework for the design of association rule mining applications and predefined templates as a means to capture the user specification of mining applications.

Although numerous research projects have been undertaken, and some are currently ongoing, Croft (1995) suggested that the current state of information extraction tools is such that it requires a considerable investment to build a new extraction application, and certain types of information are very difficult to identify. However, Croft further commented that extraction of simple categories of information is practical and can be

an important part of a text-based information system. This article highlights some potential areas of application of template mining in a digital library environment.

Use of Templates in Web Search Engines

Gaizauskas and Wilks (1998) suggested that there is a contrast between the aims of information extraction and information retrieval systems in the sense that IR retrieves relevant documents from collections, while IE extracts relevant information from documents. However, the two are complementary, and their use in combination has the potential to create powerful new tools in text processing and retrieval. Indeed, IE and IR are equally important in the electronic information environment, particularly the World Wide Web, and templates have been used both for IR and IE. Many applications of template mining mentioned above handle digital texts available on the Web, while search engines use templates to facilitate IR.

Search engines are one of the most essential tools on the Internet—they help find Web sites relating to a particular subject or topic. Search engines are basically huge databases containing millions of records that include the URL of a particular Web page along with information relating to the content of the Web page supplied in the HTML by the author. A search engine obtains this information via a submission from the author or by the search engine doing a "crawl" using "robot crawlers" of the Internet for information. The most popular search engines include: AltaVista, Excite, Hotbot, Infoseek, Lycos, Webcrawler, Yahoo, and so on.

Some search engines use templates to help end-users submit natural language queries used by search engines to conduct a search on specific topics. Two small sets of tests were conducted to see how this is done in a large search engine—AltaVista—and in a meta search engine—Ask Jeeves. The following section shows how these search engines use templates for natural language query formulation in their interfaces.

Use of Templates in Alta Vista

The Alta Vista search engine (http://www.altavista.com) helps users find information on the Web. One interesting feature of this search engine is that a user can enter one or more search terms/phrases or can type a natural language statement such as "What is the capital of Alaska?" or "Where can I find quotations by Ingmar Bergman?" Taking the second option, a simple query statement, "Where can I find information on Web search engines?" was typed in the specified box of the Alta Vista search interface (see Figure 1). Along with the results, Alta Vista came up with two templates that contain natural language sentences related to the search topic (see Figure 2). By clicking on the box at the end of the statement "How do I (Internet skill)?" the system shows a box containing various

options (Figure 3), any of which can be chosen to complete the sentence, the default one being "search through ALL web sites." By choosing this, or any other option from the box, a user can formulate a sentence-like query such as: "How do I search through all Web sites?" or "How do I learn HTML?" or "How do I use the Internet as a telephone?" and so on.

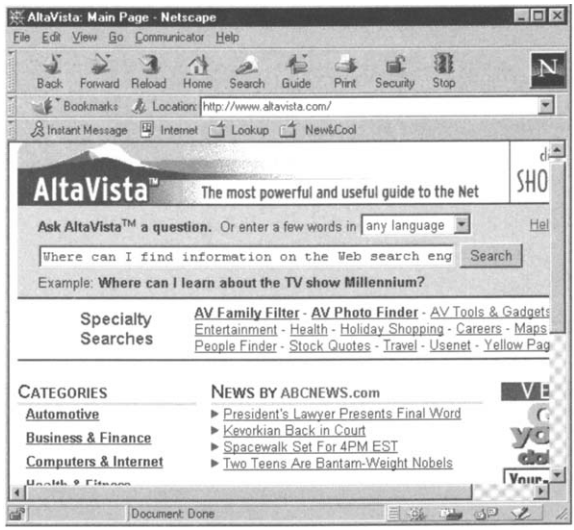


Figure 1. The Search Interface of Alta Vista.

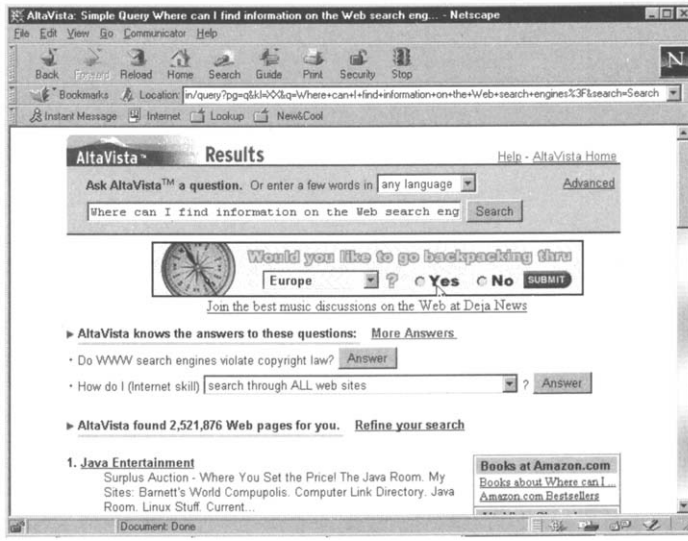


Figure 2. Output of a Simple Search.

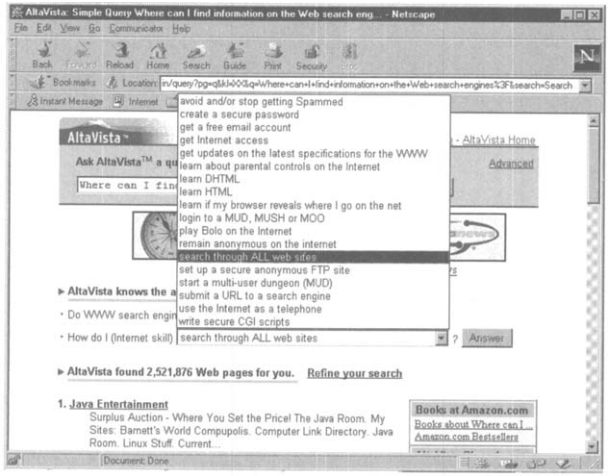


Figure 3. Options for the "(Internet skill)" Statement Slot of the Template in Figure 2.

The above examples show that the search engine uses templates and various options for the "(Internet skill)" slot in the template "How do I (Internet skill)." By clicking on the "More answers" option (see Figure 2), a user can get more such templates (see Figure 4). It may be noted that for many slots, such as "computing term," "Internet term," "search engine," and so on (see Figure 4), there are various options that can be displayed by clicking on the appropriate box. Some of these options are shown in Figure 5. Thus, the search engine uses various templates and provides options to fill in the slots to prepare sentence-like queries. Once a user prepares a search sentence by choosing the appropriate option from the drop-down box and clicks on the "Answer" button, the system conducts a search and fetches the relevant hits. However, it may be noted that the format of the query templates, and quite obviously the contents of the box showing the various options for the slots, vary from query to query. For example, when the system was asked "When will the next World Cup Football Games be held?" the templates that came up on the output screen were different (see Figures 6, 7, and 8). However, the system does not always come up with natural language query templates. For example, when the query, "What is a hurricane?" was given, the system simply produced a list of hits and no templates (see figure 9).

Use of Templates in Ask Jeeves

Ask Jeeves (http://www.askjeeves.com) is a meta-search engine that represents a model of an application using knowledge management techniques in order to better organize disparate information sources

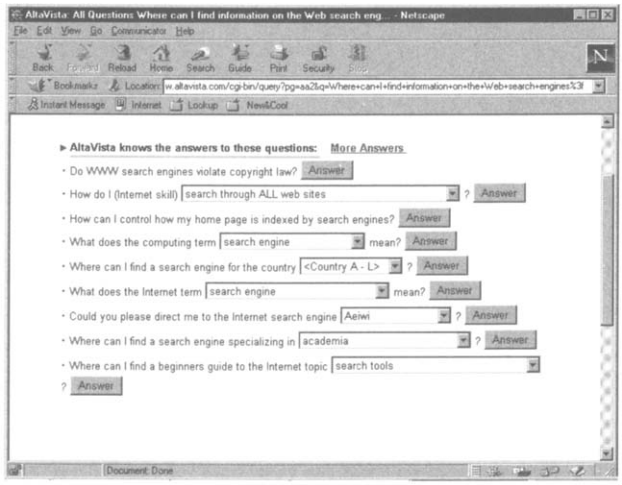


Figure 4. More Templates for the Query Shown in Figure 1.

a.

b.

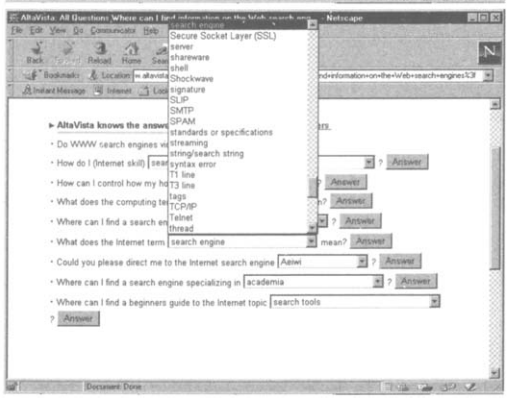


Figure 5. Options for the "computing term" (a) and "Internet term" (b) Slots in the Templates (in Figure 4).

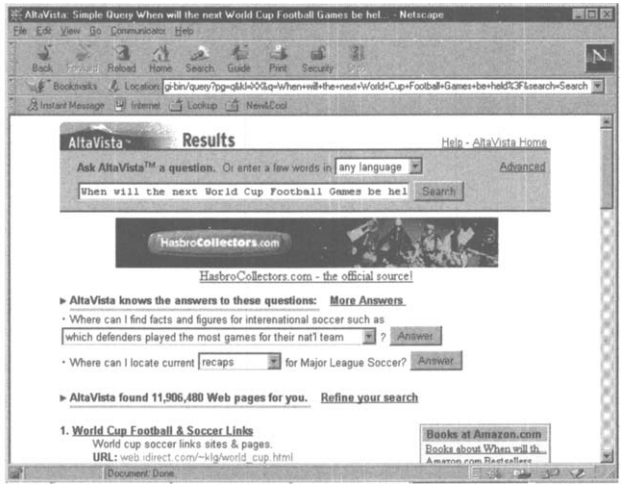


Figure 6. Search Results of the Query "When Will the Next World Cup Football Games Take Place?"

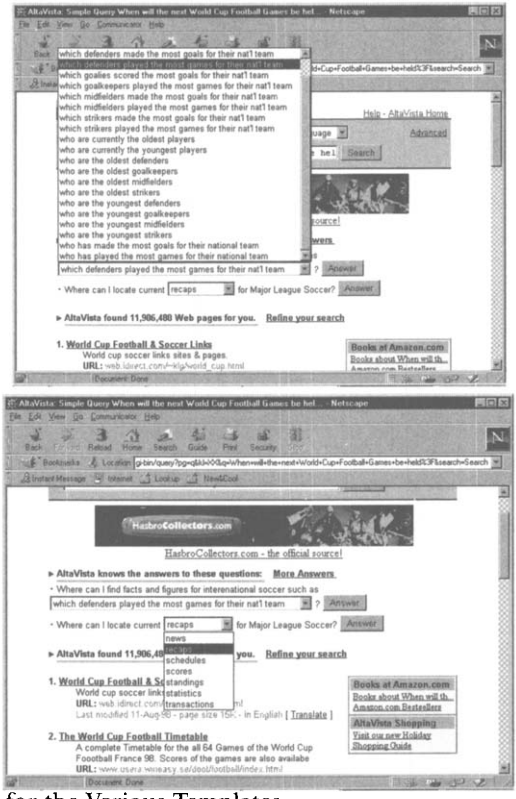


Figure 7. Options for the Various Templates.

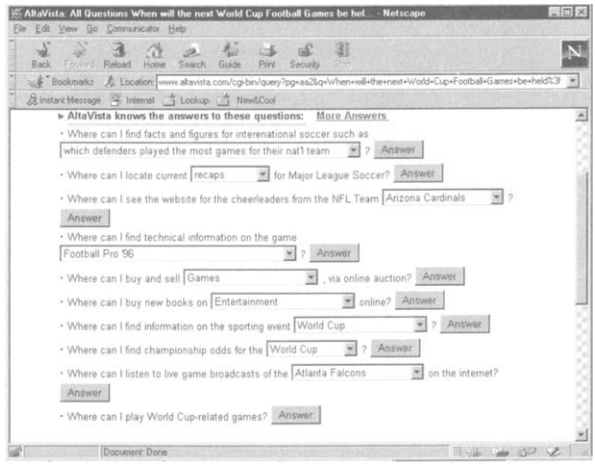


Figure 8. More Templates for the Query Shown in Figure 6.

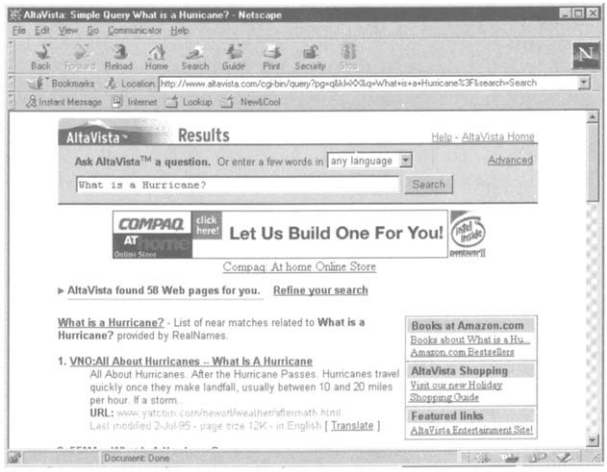


Figure 9. Output of the Query Shown in Figure 6 "What is a Hurricane?"

(Stanley, 1998). It draws on the expertise of experienced human Web searchers and encapsulates this expertise in a database so that it can be put to use by others. Various questions and their answers are manually selected by human editors who scan resources on the Web on a daily basis to build up a knowledge base of information about sites which might be used to answer common questions. The questions and the Web pages which answer them are then stored as a series of templates in the Ask Jeeves knowledge base, and keywords and concepts in a search string are matched against them in order to retrieve the questions and their corresponding Web sites (Stanley, 1998). Ask Jeeves was asked a simple question: "What is a hurricane?" (see Figure 10), and the system created a number of templates as shown in Figure 11.

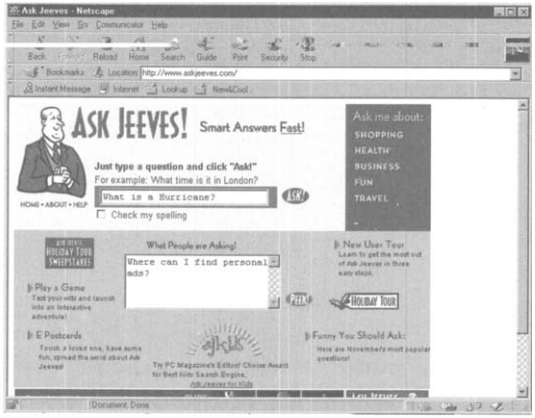


Figure 10. Search Interface of Ask Jeeves.

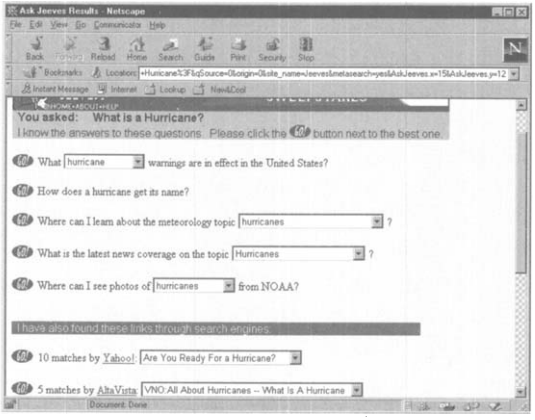


Figure 11. Output (Showing Templates) of the Query Shown in Figure 10.

Thus, the search engines use templates, though not for IE but for IR, more specifically to help users formulate a natural language query. However, the way these templates are created is quite interesting. Human experts conduct searches on the topics and, based on the search results, organize them into different groups. These groupings could be on different topics such as "computing terms," "Internet terms," "Internet skills," and so on. These are then created as slots in sentence-like queries, and the various options for the slots (each "Internet term," "Internet skill," and so on) are then presented in boxes for the end-user to select while searching.

Template Mining in Digital Libraries

The British Library DL (digital library) Program (The British Library ..., 1997) defines digital library as the widely accepted descriptor for the use of digital technologies to acquire, store, conserve, and provide access to information and materials in whatever form it was originally published. The Stanford digital library working paper (Reich & Winograd, 1995) defines a digital library as a coordinated collection of services that are based on collections of materials, some of which may not be directly under the control of the organization providing a service in which they play a role. The contents of a digital library, being digital information sources, provide ample opportunities for applying template mining resulting in the extraction of valuable information in a number of areas in a digital library environment. Four areas of such an application have been identified and how this is done is discussed in the following sections.

Automatic Creation of Citation Databases of Digital Documents

Recently there have been some works on citation studies in the Web environment. Almind and Ingwersen (1998) introduced the concept of "Webometrics"—i.e., the application of informetric methods to the WWW. They have argued that citation analysis on the WWW has not been tested in practice. In another publication, Ingwersen and Hjortgaard (1997) discussed the advantages and disadvantages of using the WWW for informetric analysis and examined the pitfalls of online informetric analysis using the ISI (Institute of Scientific Information) files.

A quick and simple examination of some issues of the electronic journal *D-Lib Magazine* (http://www.dilib.org/dilib) revealed that template mining can be used to develop citation databases automatically from the online articles. Such databases may contain information somewhat similar to the ISI databases, such as the citing author, address of the citing author, title of the citing article, keywords, and so on as well as the authors, titles, and bibliographic details of the cited articles. The simple template mining approach may be used to extract the information for

each of these fields in the citation database that can later be used for various citation analysis and other purposes. Analysis of the articles (they are called stories in *D-Lib Magazine*) published in the nine 1998 issues of *D-Lib Magazine* (January, February, March, April, May, June, July/August, September, and October) revealed a general structure of the articles (see Figure 12).

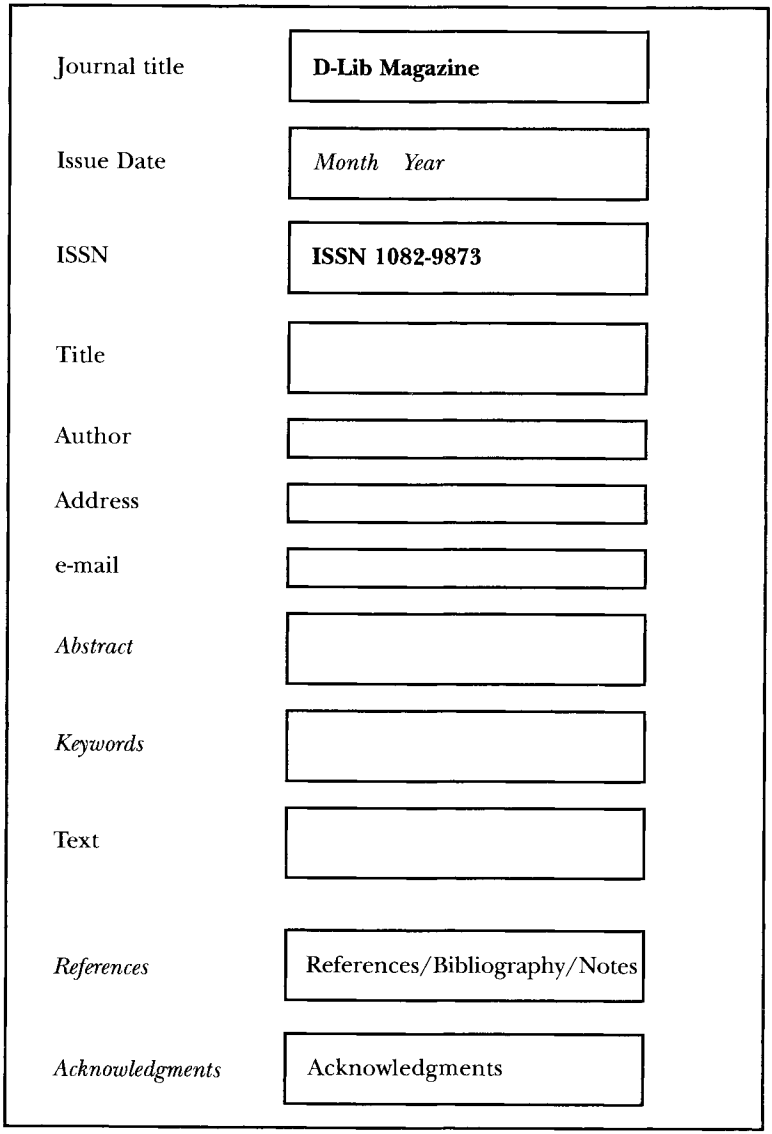


Figure 12. General Structure of Articles (called Stories) in *D-Lib Magazine.*

Figure 12 provides a general idea of the different kinds of templates that can be generated to extract information from the various sections of each article. Text that appears in bold in Figure 12 shows that the concerned text is constant—i.e., it appears in each article. Similarly, empty boxes indicate that texts appear there to indicate value for the given slot—e.g., title, author's name, and so on. Texts in some slots appear in a specific format, for example, in the "Issue Date" slot, the particular issue appears in the format "Month Year"—e.g., "February 1998." In some slots, the heading varies. For example, the heading used for references is usually "References" but the headings "Bibliography" or "Notes" are also used. This preliminary study has shown that, although this is the general structure of the articles in *D-Lib Magazine*, some articles may not have some of the slots—i.e, "Keywords," "Abstract," "References," or "Acknowledgments."

The values for some of the slots remain constant while, for most of the slots, they vary from one article to the other. For example, in the "Author" and "Address" slots, different patterns have been noticed. Articles may be written by only one author, by two authors, or by more than two authors. Again, when more than one author is involved, they may have the same or different addresses. Articles also differ in terms of layout for writing the authors' names and addresses: while in most cases they appear vertically, one after the other, in some cases they appear horizontally, one after the other. The general structure of these slots is shown in Figure 13.

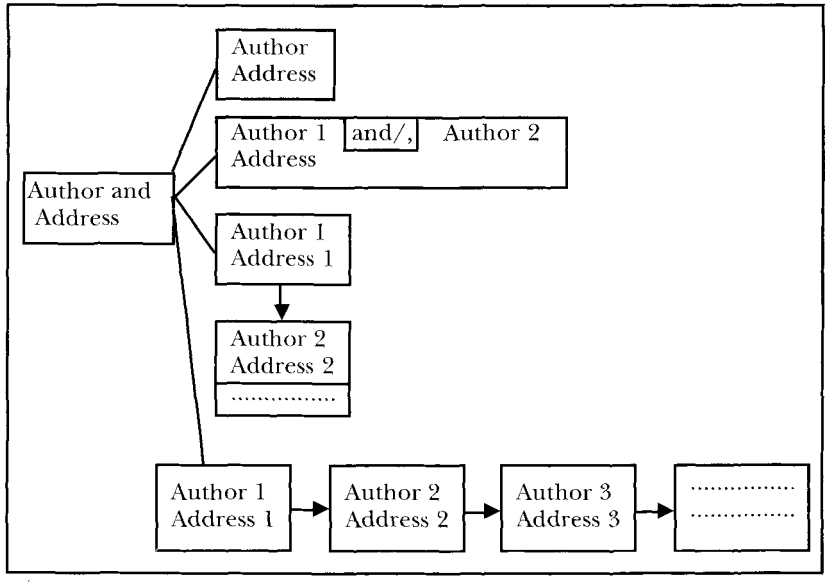
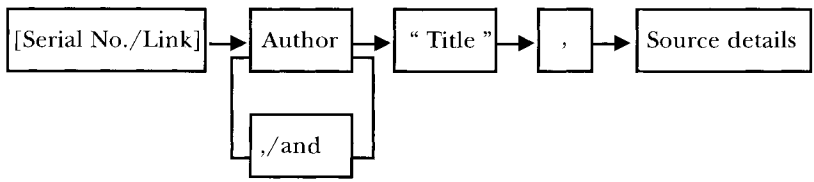


Figure 13. Template for the Author and Address Information.

Similarly, the contents of the "References" slot vary significantly depending on the type of item referred to. Some articles have only print versions while others are available only in electronic form. Therefore, the citations relating to the sources vary: there may be journal title, volume number, and so on or there may only be the URL or both the bibliographic details and the URL. Although the general pattern of citations as they appear at the end of each article appears to be quite simple (see Figure 14), and therefore amenable to template mining, a closer look at the references indicates that there are a number of irregularities there. For example, different authors use different citation styles and, sometimes, within the same article, authors follow different citation styles for similar types of material (see Figure 15). If these irregularities are sorted out, the template matching technique can be used to extract relevant information, including URLs, from the citations. There are two ways to sort out these irregularities and to ensure a standard citation style. The first one could be to impose rigorous editorial practice, but that would be more expensive and time consuming, causing delays in publication. The second, and less expensive, approach may be to prepare templates for each type of citation and make them available online. Authors can download and make use of these templates for preparing the list of references. This will ensure a standard citation style. The same practice may be followed for the other parts of the article, and eventually the whole structure of the articles can be standardized, thereby facilitating the use of template mining.



Sample References from D-Lib Magazine

[Chen et al., 1996] Hsinchun Chen, Chris Schuffels, and Rich Orwig, "Internet Categorization and Search: A Machine Learning Approach," Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, Volume 7, Number 1, Pages 88-102, 1996.

[5] Harnad E., Print Archive and Psycoloquy and BBS Journal Archives <http://www.princeton.edu/~harnad>

[11] Trudi Bellardo Hahn, Text Retrieval Online: Historical Perspective on Web Search Engines, pp. 7-10, Bulletin of the American Society for Information Science, April/May, 1998.

Figure 14. Simple Structure of References in the *D-Lib Magazine.*

[ieeePC97]
Bacon J, Bates J and Halls D, Location oriented multimedia, IEEE Personal Communications 4(5), pp 48-57, October 1997.

[Phelps and Wilensky, 1996b] Thomas A. Phelps and Robert Wilensky, "Multivalent Documents: Inducing Structure and Behaviors in Online Digital Documents", Proceedings of Hawaii International Conference on System Sciences '96 (Best Paper Award, Digital Documents Track).

[Salton, 1989]Gerard Salton, Automatic Text Processing, Addison-Wesley, 1989.

[Weiss et al., 1996]Ron Weiss, David Gifford et al., "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering", Proceedings of the Seventh ACM Conference on Hypertext, March 1996, Washington, DC.

[Weiss and Indurkhya, 1993] S. Weiss and N. Indurkhya, "Optimized rule induction," IEEE Expert 8, 6, 61-69.

[Yahoo]Yahoo [http://www.yahoo.com/]

[Zhu et al., 1997] Quan Zhu et al., "Searching for Parts and Services on the Web", Proceedings of International Symposium on Research, Development, and Practice in Digital Libraries, Nov. 18 - 21, 1997, Tsukuba, Japan.

[1] Tom Sanville of the OhioLINK consortium has noted in public presentations a correlation between ease of access and use and the amount of use.

[2] TULIP: The University Licensing Program <http://www.elsevier.nl/locate/tulip>

[3] Red Sage: Final Report <http://www.springer-ny.com/press/redsage>

[5] Harnad E. Print Archive and Psycoloquy and BBS Journal Archives <http://www.princeton.edu/~harnad>

[9] JSTOR http://www.jstor.org

[Arms 1995] R. Kahn and R. Wilensky, A Framework for Distributed Digital Object Services, (May 1995).

[Bearman 1998] D. Bearman and J. Trant, Authenticity of Digital Resources:Towards a Statement of Requirements in the Research Process, D-LIB on-line magazine, (June 1998).

C. Lynch et al., A White Paper on Authentication and Access Management Issues in Cross-organizational Use of Networked Information Resources,

(Note: The underlines in the last three references indicate that they are hyperlinked to the respective URLs)

Figure 15. Sample References from Different Issues of *D-Lib Magazine* Showing the Lack of Standards in the Citation Style.

Automatic Extraction of Information from News Items in Electronic Journals

Electronic journals often contain news items, and important information can be extracted for the creation of databases or for any other use by a simple template mining approach. A scan through the pages of the nine issues of *D-Lib Magazine* revealed that there is a section called "Goings on" that provides information on conferences/seminars/workshops, and so on under the heading "Clips & Pointers." A scan through the items appearing under the heading "Goings on" in all the 1998 issues of *D-Lib Magazine* revealed that the seminar/conference/workshop announcements follow a general pattern as shown in Figure 16. This shows that a simple template can extract information about the various forthcoming seminars, conferences, and so on. Particular information, such as the place, date, Web address, and so on, can also be extracted by such templates. However, once again such a template mining approach calls for a standard format and layout. Templates can also be generated to extract further information such as specific topics, contact address, deadlines, and so on.

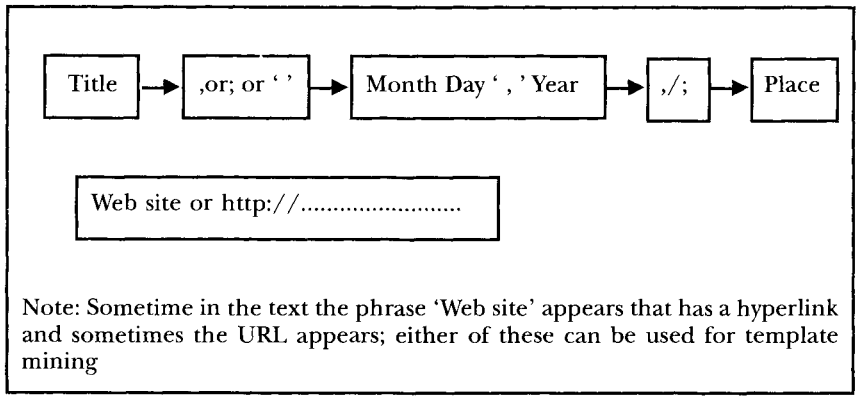


Figure 16. Simple Templates for Conference/Seminar/Workshop Announcements.

Automatic Identification of Funding/Sponsoring Agencies for Research

A scan through the "Acknowledgment" section in the articles revealed that they contain information about the funding/sponsoring agency's name, address, grant number, and so on. A study of the various articles appearing in the electronic journals can help generate a pattern, and thereby appropriate templates, that will be able to extract the relevant information for further use. Again, in order to standardize the practice

of providing these items of information, the editorial board of the journal can make appropriate templates available to the authors.

Information Extraction Using Metadata and Template Mining

Metadata are data about data. However, this definition is too simple and does not tell us their characteristic features. A better definition has been provided by Dempsey and Heery (1998) according to whom metadata are data associated with objects which relieve their potential users of having to have full advance knowledge of the existence of characteristics. Younger (1997) defined metadata as documentation about documents and objects; they describe resources, indicate where they are located, and outline what is required in order to use them successfully. There are several metadata schemes created by library and information professionals over the years, the most prominent ones being the MARC formats, the AACR2 catalog formats, subject headings lists (such as the LCSH), and classification schemes such as LC, DDC, UDC, and so on. Each of these schemes is constructed by experts in the field from an understanding of the specific domains, information resource needs, and the requirements for describing documents. While these schemes have been used for bibliographic access and control for decades, there remains the question of how to catalog and index materials available on the Internet using these schemes. This has given rise to a thought that electronic documents need to be *self-indexed* (as opposed to the assignment of cataloging and indexing tags and value added by cataloging and indexing agencies or library staff). However, it is obvious that, in order for the documents to be self-indexed, a core set of metadata elements must be identified, and each creator of electronic documents should be able to implement it in the record that he creates. With this objective, a simple resource description set of data has emerged—the Dublin Core (http://purl.org/metadata/dublin_core). The Dublin Core metadata set prescribes fifteen elements, namely (http://www.lub.lu.se/cgi-bin/nmdc.pl): title, creator, subject (keywords, controlled vocabulary, and classification), description (abstract and content description), publisher, contributor (other than the creator), date, type (category of the resource), format (HTML, Postscript, etc.), identifier (URL, string or number used to identify the resource), source (from which this resource is derived), language, relation (with other resources), coverage (spatial and/or temporal characteristics of the resources), and rights (link to a copyright notice, etc.).

Weibel (1995,1996) suggested that, in order to enable information creators to apply metadata, a mechanism for embedding the data within HTML documents had to be established. The Dublin Core looks at one aspect of metadata—i.e., the simple description—but, as Dempsey and Heery (1998) suggested, there is a need for more complex description for

particular specialist domains. In 1996, there was a conference organized by UKOLN and OCLC to examine the various metadata issues including the Dublin Core. This meeting gave rise to a proposal, the Warwick Framework proposal (Lagoze, 1996) (named after the place of the conference), calling for an architecture for the interchange of various metadata packages.

Metadata are an important tool for resource discovery from digital documents. They not only help users locate the required information resources, they also help in the examination and selection (or rejection) of the retrieved items. Various fields, such as subject, description, language, sources, and so on, can provide necessary information for examining the relevance of the retrieved resources. Web search engines, as an aid during the examination phase, generally construct an information surrogate to display the search hits. These surrogates generally consist of the URL, title of the Web page, and some summary text that is derived with the aid of some heuristics (Lagoze, 1997).

Lagoze (1997) argued that, recognizing the limitations of the current search engines, researchers are now actively pursuing both standards for descriptive surrogates for networked objects and methods for associating surrogates with those objects. Such studies aim at developing a number of surrogate templates that would facilitate the resource discovery process. The template mining approach can be used effectively for the resource discovery process. The metadata tags, such as subject and description in the Dublin Core metadata format, for example, can be extended and be made more structured by specifying various templates. The following section briefly describes how this can be achieved.

Use of the Principles of Pre-Coordinate Indexing Systems in Building Subject Templates

Information retrieval systems have used two different types of indexing systems, namely, pre-coordinate indexing and post-coordinate indexing (for detailed discussions, see Lancaster, 1998; Foskett, 1996). Pre-coordinate indexing systems—classification schemes like the Dewey Decimal Classification (DDC) (Dewey, 1996), Universal Decimal Classification (UDC, 1985), and Colon Classification (CC) (Ranganathan, 1965), and so on that use artificial notations or subject indexing systems such as relational indexing (Farradane, 1980a, 1980b), PRECIS (Austin & Dykstra, 1984), and POPSI (Bhattacharyya, 1981) that use natural language terms—represent the content of an information resource by synthesizing the various components of the subject and organizing these in a specific order. On the contrary, post-coordinate indexing systems do not rely on the a priori relations or organization of the constituent search terms; rather, retrieval is performed by searching through each individual term and then the output is generated based on the coordination of the terms at the

retrieval stage; for example, according to the principles of set theory. Most of the modern day IR systems, including the search engines, follow the principles of post-coordinate indexing systems. However, it is proposed that, for resource discovery from digital libraries or from the Web, one can use the principles of pre-coordinate indexing systems. The following paragraph suggests a simple approach to this.

The basic tenet of this approach is that pre-coordinate indexing systems help indexers generate subject index entries that represent the subject matter of the document concerned. This approach has been successfully used in libraries for organizing library materials on the shelves based on classification numbers. These class numbers are created according to the principles of pre-coordinate indexing systems, and the notations represent the content of the document concerned. Similarly, this approach has been used in preparing alphabetical subject index entries for documents in national bibliographies and in other bibliographic databases. One major drawback of these pre-coordinate indexing systems is that it is largely human dependent because human indexers need to analyze the documents and prepare the subject statements, which are then manipulated by computers for generating multiple entries or are used to prepare the class number. In other words, this process involves a significant amount of human expertise and time and therefore is a slow and expensive process. As a result, adopting this approach is almost impossible with a large collection and is impossible in the Web environment with millions of documents.

The aforementioned problem could be solved if this task was accomplished by the author or the generator of information resources. If authors can somehow indicate the key concepts treated in the documents, along with an indication of the appropriate categories where they belong, then subject statements or simple surrogates can be prepared automatically. A template can be provided to the generator of information resources—e.g., an author of an article or a report may fill in the various slots with appropriate information. Such templates can be generated using the various categories proposed in pre-coordinate indexing systems—e.g., the fundamental categories of Ranganathan (1967), the nine relational categories proposed by Farradane (see for discussion, Farradane, 1980a, 1980b; Chowdhury, 1989), the various role operators proposed in PRECIS by Austin (Austin & Dykstra, 1984), the various categories proposed by Bhattacharyya (1981) in POPSI, and so on. The author or the generator of the information resource should be able to understand the connotation of each category and thus should be able to fill in the template according to the semantic content of the concerned information resource. For example, the author of a document entitled "Internet as a Tool for Management of OPACs in the Libraries in Singapore" is given a simple template created according to the PRECIS Role Operators (see Austin & Dykstra, 1984; Chowdhury, 1995). Now the author can fill in the

slots in the template according to the role of each term in the given document as shown in Figure 17. Note that the template shown in Figure 17 is not the full implementation of PRECIS and therefore does not show all PRECIS operators; this is rather indicative of the kind of application that one can build according to the principles of any pre-coordinate indexing system. Which system is the best, and yet easy for an author to understand and apply, is a matter for further research.

Location: Singapore

Key system/Object of transitive action/Agent of intransitive action/Effect of action: Libraries
Part/property: OPAC

Action, Discipline, etc.: Management

Agent/Performer of transitive action/Intake/Factor: Internet

Viewpoint:

Selected Instance, study region, sample population:

Form of Document, target user:

Figure 17. Simple Templates according to the PRECIS Role Operators.

Such templates, prepared according to the principles of any pre-coordinate indexing system or a modified version of that, can be used both for better retrieval and for the preparation of document surrogates. Better retrieval can be achieved because of the semantic values attached to the terms; this would help reduce the *false drops*. Document surrogates can be prepared automatically according to the prescribed rules of the concerned indexing system—e.g., according to the principle of generating the index entries in PRECIS (Austin & Dykstra, 1984).

CONCLUSION

The explosive growth in our capabilities to collect and store data over the past decades has given rise to a new field of study, called knowledge

discovery in databases, that is concerned with the creation of a new generation of tools and techniques for automated and intelligent database analysis. Raghavan et al. (1998) suggest that KDD refers to the whole process in the path from data to knowledge and to use descriptive phrases for specific tasks in the process, such as pattern extraction methods, pattern evaluation methods, or data cleaning methods. Thus, simply speaking, KDD is the process of deriving useful knowledge from real-world databases through the application of pattern extraction techniques. The grand challenge of KDD is, therefore, to automatically process large quantities of raw data, identify the most significant and meaningful patterns, and present these as knowledge appropriate for achieving a user's goals.

This discussion has proposed that template mining, which is based on pattern recognition and pattern matching in natural language texts, can be used for extracting different kinds of information from digital documents or text databases. Citation databases can be built automatically based on the template mining approach, from digital documents, such as from articles in electronic journals. Template mining can also be used to extract different types of information, such as information about the funding/sponsoring agencies of research projects as they appear in the acknowledgments section of articles. Another application could be to identify inter-document links by tracing the hypertext links using the "http://...." template, and thus a network of articles in a specific domain can be built that would be useful for researchers in the subject concerned. This article has also indicated that template mining can be used to extract various items of news from digital documents such as in the extraction of conference information from electronic journals, and so on.

This article has also indicated that the template mining approach can be incorporated within the metadata format in order to facilitate better information retrieval and to enable automatic generation of document surrogates that would help the end-user filter the search output. This is very necessary, particularly in the WWW environment where a given search may retrieve several thousands, even millions, of records. This could be achieved by using the concept categorization and organization principles of the pre-coordinate indexing systems. Such pre-coordinate systems will improve the quality of the output of search engines that basically follow the principles of post-coordinate indexing systems. In other words, while post-coordinate indexing principles used in the search engines will retrieve digital information resources, the principles of pre-coordinate indexing systems may be used to filter them.

However, all the above-mentioned applications will be possible provided the digital information resources appear in standard format and layout. The first few applications of template mining mentioned above will require strict adherence to formats. This should not be too difficult, as authors are used to following author instructions issued by publishers/

editors of printed journals and publication houses. It should not be a big problem to implement the same in the electronic environment. The only requirement will be to formulate appropriate guidelines and, if possible, appropriate templates can be made available to the authors that could be used online while preparing the documents. For example, simple templates may be prepared for each type of publication, and authors may be required to just click on the type of document, such as journal article, conference paper, online sources, and so on, to get the appropriate slots to fill-in with data for author, title, source details, and so on. This would not only ensure consistency in the references but would also facilitate template mining applications.

For the last application proposed above, much more work needs to be done. It may be incorporated as an element in the metadata format, or may be added as a required field in any digital document, or embedded in HTML. However, some experiments must be conducted in order to determine which pre-coordinate indexing system will be more appropriate and yet easy for the authors to use while preparing the digital information sources. This may be kept at a simple level just by creating slots for each category of the chosen pre-coordinate indexing system or may be made more complex by incorporating the phase relations proposed by Ranganathan (1967, 1987) as well. However, with the increasing complexity of the system, it may be more difficult, and therefore more inhibiting, for the authors who are required to fill in the slots in the templates. Nevertheless, this job has to be done by the generators of digital information resources, otherwise it may be too difficult and expensive for any agency to analyze each digital document and fill in the slots as done in traditional libraries for classifying and indexing materials. Initially, it may seem an extra burden to the authors but, as we need to follow instructions from editors and publishers and follow HTML and similar standards while preparing hard copy and/or digital documents, we may have to do this in order to make our generated information more widely and easily available to potential users. Eventually, all these additional activities may be incorporated within popular software and, as we can now use editors for creating HTML documents rather than coding them by hand, authors may just fill in the templates as part of their document creation task. Detailed experiments are currently underway and, upon successful completion of this research, we expect to develop a model for template mining from digital information resources.

ACKNOWLEDGMENT

The author gratefully acknowledges the encouragement and support obtained while writing this article from Michael F. Lynch, Professor Emeritus, Department of Information Studies, University of Sheffield, Sheffield, England.

REFERENCES

Ai, C. S.; Blower, P. E.; & Ledwith, R. H. (1990). Extraction of chemical reaction information from primary journal text. *Journal of Chemical Information and Computer Sciences, 30*(2), 163-169.

Almind, T. C., & Ingwersen, P. (1998). Informetric analyses on the World Wide Web: Methodological approaches to "Webmetrics." *Journal of Documentation, 54*(4), 404-426.

Andersen, P. M.; Hayes, P. J.; Huettner, A. K.; Schmandt, L. M.; Nirenburg, I. B.; & Weinstein, S. P. (1992). Automatic extraction of facts from press releases to generate news stories. In *Third Conference on Applied Natural Language Processing* (31 March-3 April, 1992, Trento, Italy) (pp. 170-177). Morristown, NJ: Association of Computational Linguistics.

Andersen, P. M., & Huettner, A. K. (1994). Knowledge engineering for the JASPER fact extraction system. *Integrated Computer-Aided Engineering, 1*(6), 473-493.

Austin, D., & Dykstra, M. (1984). *PRECIS: A manual of concept analysis and subject indexing* (2d ed.). London, England: British Library.

Baralis, E., & Psaila, G. (1997). Designing templates for mining association rules. *Journal of Intelligent Information Systems, 9*(1), 7-32.

Bhattacharyya G. (1981). Some significant results of current classification research in India. *International Forum on Information and Documentation, 6*(1), 11-18.

The British Library Research and Innovation Centre. (1998). *The British Library Digital Library Programme.* Retrieved December 16, 1998 from the World Wide Web: http://www.bl.uk/services/ric/diglib/digilib.html.

Chong, W., & Goh, A. (1997). FIES: Financial information extraction system. *Information Services & Use, 17*(4), 215-223.

Chowdhury, G. G. (1989). *Nature and applicability of Farradane's relational analysis with particular reference to the preparation of linear index entries.* Unpublished doctoral dissertation, Jadavpur University, Calcutta, India.

Chowdhury, G. G. (1992). *Application of natural language processing to chemical patents.* Unpublished doctoral dissertation, University of Sheffield, Sheffield, England.

Chowdhury, G.G. (1995). *PRECIS: A workbook.* Calcutta, India: IASLIC.

Chowdhury, G. G., & Lynch, M. F. (1992a). Automatic interpretation of the texts of chemical patent abstracts. Part 1: Lexical analysis and categorization. *Journal of Chemical Information and Computer Sciences, 32*(5), 463-467.

Chowdhury, G. G., & Lynch, M. F. (1992b). Automatic interpretation of the texts of chemical patent abstracts. Part 2: Processing and results. *Journal of Chemical Information and Computer Sciences, 32*(5), 468-473.

Coates-Stephens, S. (1992). *The analysis and acquisition of proper names for robust text understanding.* Unpublished doctoral dissertation, City University, London, England.

Costantino, M.; Morgan, R. G.; & Collingham, R. J. (1996). Financial information extraction using pre-defined user-definable templates in the LOLITA system. *Journal of Computing & Information Technology, 4*(4), 241-255.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Croft, W. B. (1995). What do people want from information retrieval?: The top 10 research issues for companies that use and sell IR systems. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/november95/11croft.html.

De Jong, G. (1982). An overview of the FRUMP system. In W. Lehnert & M. H. Ringle (Eds.), *Strategies for natural language processing* (pp. 149-176). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dempsey, L., & Heery, R. (1998). Metadata: A current view of practice and issues. *Journal of Documentation, 54*(2), 145-172.

Dewey, M. (1996). *Dewey decimal classification and relative index.* Albany, NY: Forest Press.

Farradane, J. E. L. (1980a). Relational indexing. Part 1. *Journal of Information Science, 1*(5), 267-276.

Farradane, J. E. L (1980b). Relational indexing. Part 2. *Journal of Information Science, 1*(6), 313-324.

Foskett, A. C. (1996). *The subject approach to information* (5th ed.). London, England: Library Association Publishing.

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: Beyond document retrieval. *Journal of Documentation, 54*(1), 70-105.

Gaizauskas, R., & Humphreys, K. (1997). Using a semantic network for information extraction. *Natural Language Engineering, 3*(2/3), 147-169.

Guarino, N. (1997). Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In M. T. Pazienza (Ed.), *Information extraction: A multidiciplinary approach to an emerging information technology international summer school* (No. 1299: Lecture Notes in Artificial Intelligence) (pp. 139-168). New York: Springer.

Ingwersen, P., & Hjortgaard C. F. (1997). Data set isolation for bibliometric online analysis of research publication: Fundamental methodological issues. *Journal of the American Society for Information Science, 48*(3), 205-217.

Jacobs, P., & Rau, L. F. (1990). SCISOR: Extracting information from on-line news. *Communications of the ACM, 33*(11), 88-97.

Jones, P. A., & Paice, C. D. (1992). A "select and generate" approach to automatic abstracting. In T. McEnery & C. D. Paice (Eds.), *Proceedings of the BCS 14th Information Retrieval Colloquium* (pp. 141-154). Berlin, Germany: Springer-Verlag.

Kemp, N. M. (1995). *The application of natural language processing to chemical patents.* Unpublished doctoral dissertation, University of Sheffield, Sheffield, England.

Lagoze, C. (1996). The Warwick Framework: A container architecture for diverse sets of metadata. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

Lagoze, C. (1997). From static to dynamic surrogates: Resource discovery in the digital age. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/june97/06lagoze.html

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice* (2d ed.). Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois.

Lawson, M.; Kemp, N.; Lynch, M. F.; & Chowdhury, G. G. (1996). Automatic extraction of citations from the text of English language patents: An example of template mining. *Journal of Information Science, 22*(6), 423-436.

Lytinen, S. L., & Gershman, A. (1986). ATRANS: Automatic processing of money transfer messages. In *AAAI '86* (The Fifth National Conference on Artificial Intelligence, August 11-15, 1986, Philadelphia, Pennsylvania) (pp. 1089-1093). Los Altos, CA: Morgan Kaufmann.

Paice, C.D. (1981). The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases. In R. N. Oddy, S. E. Robertson, C. J. Van Rijsbergen, & P. W. Williams (Eds.), *Information retrieval research* (pp. 172-191). London, England: Butterworths.

Paice, C. D., & Husk, G. D. (1987). Towards the automatic recognition of anaphoric features in English text: The impersonal pronoun "it." *Computer Speech and Language, 2*(2), 109-132.

Postma, G. J., & Kateman, G. (1993). A systematic representation of analytical chemical actions. *Journal of Chemical Information and Computer Sciences, 33*(3), 350-368.

Postma, G. J.; van der Linden, J. R.; Smits, J. R. M.; & Kateman, G. (1990a). TICA: A system for the extraction of data from analytical chemical texts. *Chemometrics & Intelligent Laboratory Systems, 9*(1), 65-74.

Postma, G. J; van der Linden, J. R.; Smits, J. R. M.; & Kateman, G. (1990b). TICA: A system of extraction of analytical chemical information from texts. In E. J. Karjalainen (Ed.), *Scientific computing and automation (Europe)* (pp. 407-414). Amsterdam: Elsevier.

Raghavan, V. V.; Deogun, J. S.; & Sever, H. (1998). Introduction (to the special issue on knowledge discovery and data mining). *Journal of the American Society for Information Science, 49*(5), 397-402.

Ranganathan, S. R. (1987). *Colon classification* (7th ed.). Bangalore, India: Sarada Ranganathan Endowment for Library Science.

Ranganathan, S. R. (1967). *Prolegomena to library classification* (3d ed.). Bangalore, India: Sarada Ranganathan Endowment for Library Science.

Reich, V., & Winograd, T. *Working assumptions about the digital library* (Stanford Digital Library Working Paper Feb. 23, 1995). Retrieved December 16, 1998 from the World Wide Web: http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1995-0006.

Sager, N. (1981). *Natural language information processing: A computer grammar of English and its applications.* Reading, MA: Addison-Wesley.

Sasaki, Y. (1998). Learning of information extraction rules using ILP-programming report. In *PADD98* (Proceedings of the Second International Conference on the Practical Application of Knowledge Discovery and Data Mining, London, 25-27 March 1998) (pp. 195-205). Blackpool, England: Practical Application.

Schank, R. C. (1975). *Conceptual information processing.* Amsterdam: North-Holland.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Lawrence Erlbaum.

Schank, R. C., & Riesbeck, C. K. (1981). *Inside computer understanding: Five programs plus miniatures.* Hillsdale, NJ: Lawrence Erlbaum.

Shuldberg, H. K; Macpherson, M.; Humphrey, P.; & Corely, J. (1993). Distilling information from text: The EDS template filler system. *Journal of the American Society for Information Science, 44*(9), 493-507.

Stanley, T. (1998). Ask Jeeves: The knowledge management search engine. *Ariadne, Vol. 17.* Retrieved December 11, 1998 from the World Wide Web: http://www.ariadne.ac.uk/issue17/search-engines/intro.html

UDC: BS 1000 International Medium Edition. English text. (1985). London, England: British Standards Institution.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Wakao, T.; Gaizauskas, R.; & Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *COLING '96* (The 16th International Conference on Computational Linguistics, August 5-9, 1996, Copenhagen, Denmark) (pp. 418-423). Copenhagen, Denmark: Center for Sprogteknologi.

Weibel, S. (1995). Metadata: The foundations of resource description. *D-lib Magazine.* Retrieved December 11, 1998 from the World Wide Web: http://www.dlib.org/dlib/July95/07Weibel.html.

Weibel, S. (1996). *A proposed convention for embedding metadata in HTML.* Retrieved December 10, 1998 from the World Wide Web: http://www.oclc.org/~weibel/html-meta.html.

Younger, J. A. (1997). Resources description in the digital age. *Library Trends, 45*(3), 462-487.

Zamora, E., & Blower, P. E. (1984a). Extraction of chemical reaction information from primary journal text using computational linguistic techniques: 1. Lexical and syntactic phases. *Journal of Chemical Information and Computer Sciences, 24*(3), 176-181.

Zamora, E., & Blower, P. E. (1984b). Extraction of chemical reaction information from primary journal text using computational linguistic techniques: 2. Semantic phase. *Journal of Chemical Information and Computer Sciences, 24*(3), 181-188.

CINDI: A Virtual Library Indexing and Discovery System

BIPIN C. DESAI, RAIJAN SHINGHAL, NADER R. SHAYAN, AND YOUQUAN ZHOU

ABSTRACT

THIS ARTICLE DESCRIBES A SYSTEM CALLED CINDI for cataloging and searching documents in a distributed virtual library. When putting a document in the library, the author provides and registers metadata in the form of a semantic header for the document. The semantic header contains information on both the syntactic and semantic content of the document. An expert system simulating the expertise of a cataloging librarian helps the provider fill the semantic header according to accepted library practice. Later, if someone is searching for documents in the library, then this searcher is helped by another component of the expert system in properly formulating the query. This component simulates the expertise of a reference librarian. The system then uses information provided by the semantic headers in locating and accessing documents wanted by the searcher.

INTRODUCTION

A virtual library is a collection of electronic documents and resources distributed across a computer communication network (Saunders, 1993). These documents must be cataloged adequately so that a future interested reader (searcher) can find and access them with relative ease. Many systems (Kahle, 1991; Pinkerton, 1994; Mauldin, 1995; Welcome, 1995) catalog a document on the basis of words selected from it. They do not use the document's semantic contents but generally use a program (called a robot, worm, spider, or crawler [Web robots, 1996]) which traverses the network accessing the documents to be cataloged.

Bipin C. Desai, Raijan Shinghal, Nader R. Shayan, and Youquan Zhou, Department of Computer Science, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, CANADA H3G 1M8
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 209-233

An efficient cataloging system calls for a precise description of the semantic contents of documents. A number of systems have addressed the problem of cataloging among which CORE (Cromwell, 1994), MARC (Byrne, 1991; Crawford, 1984; Petersen & Molholt, 1990), MLC (Horny, 1985; Ross & West, 1985; Rhee, 1985), and TEI (Gaynor, 1994; Giordano, 1994) can be mentioned. These systems, however, are mainly designed for professional catalogers. Creating indexes based on search robots has the following disadvantages: repeated attempts by robots to find new resources would increase the traffic on the network; the number of these robots is increasing and system administrators would likely disallow visits by robots; a robot-based approach would become difficult to justify if the network switches to a fee-for-use mode of operation (Brody, 1995; Brownlee, 1995; Cocchi, Estrin, Shenker, & Zheng, 1991; MacKie-Mason, 1997). Searching with the more recent indexing systems (AltaVista, InfoSeek, Lycos, Yahoo) is cumbersome since the number of hits can be prohibitive due to poor selectivity of the supported search terms (Desai, 1997a).

Metadata should be designed so as to provide the semantic content of an information resource and be better suited to support its subsequent discovery than the resource itself. In many cases, the resource itself may not be able to provide its semantic contents by its nature, or it may do so only after a fairly extensive and time-consuming computation. Examples of such resources are the following forms of information: audio, video, and collections of program codes. Our metadata takes the form of a *semantic header* (SH) (Desai, 1994a). Details of SH and its comparison to the Dublin Metadata Element List (DMEL) are described by Desai (1997). The use of the DMEL in representing Web objects is given by Qin (1998).

When an author puts a document on the net, she is the one who knows the document well and can semantically describe it best. Accordingly, she fills in the slots in the semantic header. For an efficient search, the index is stored in database registries distributed across the network. Since the document provider fills her own semantic header, costly professional indexing is not required.

In this article, we describe an indexing and discovery system called CINDI (Concordia INdexing and DIscovery System), which helps a document provider fill in the semantic header for her document and register it on the net (see Figure 1). Once registered, CINDI provides the facility for a searcher to locate the semantic header and then the document. CINDI thus allows a document to be searched not only on its syntax but also on its semantics. In this article, we use the term "provider" for one who makes a document available on the Internet; a "searcher" is one who looks for document(s); a "user" can be a provider or a searcher.

The organization of this article includes a discussion of the knowledge discovery problem; an overview of CINDI; the registering and maintenance of the semantic header; the expert and database system used;

and the communication process. Owing to space limitation, the last two sections describe briefly the searching and annotation features of CINDI. The current implementation status of CINDI and our future plans are given in the conclusion.

Discovery on the Internet

In June 1995, we made a series of tests on a number of then existing Internet indexing systems; these were ALIWEB, DACLOD, EINet Galaxy, GNA Meta-Library, Harvest, InfoSeek, Lycos, Nikos, RBSE, World Wide Web Catalog, WebCrawler, WWW, and Yahoo. The intent of these tests was to determine how many URLs to documents containing the target search strings Bipin (AND) Desai were indexed by these systems. The results obtained are given in Table 1 which shows the number of hits, mis-hits, and misses (Desai, 1995a).

In this table and the following tables, the number of hits is the count of the documents found to be relevant to the query. The number of duplicates is the number of times the same document was retrieved by the indexing system using different components of the search criteria or when the same document is being served from more than one site. In the more recent search engines, the systems tend to eliminate the former form of duplicates; however, the same document accessible from more than one site is replicated in the result. The number of mis-hits is that of irrelevant documents, and the number of misses is the number of relevant documents missed by the search system.

Many of these pioneering indexing systems, existing in mid 1995, are no longer active. In the meantime, a number of new systems, such as Alta Vista, OpenText, Hotbot, and so on have emerged. Many workers in the domain of the digital virtual library feel that these newer systems have addressed many of the issues we raised in designing the CINDI System.

Table 1.
Search Statistics for Using the Search Term Bipin (AND) Desai: June 1995

Search System	*Number of Hits*	*Number of Duplicates*	*Number of Mis-hits*	*Number of Items Missed*
Aliweb	none	-	-	25
DACLOD	none	-	-	25
EINet	6	0	4	23
GNA Meta Lib.	none	-	-	25
Harvest	none	-	-	25
InfoSeek	7	0	0	18
Lycos	231	2	222	18
Nikos	none	-	-	25
RBSE	8	-	8	25
W3 Catalog	none	-	-	25
Web Crawler	7	3	0	21
WWW	2	0	0	23
Yahoo	none	-	-	25

The next series of tests was done from September through October 1997 to find the number of relevant documents that could be located by the then current search engines and to evaluate the usefulness of the index entries retrieved. Relevance of a document could be judged easily once the target set was known. We repeated the test performed in 1995 with the same search words. At the time of the test, some 325 URLs were known to contain the words "Bipin" and "Desai." These represent Web documents pertaining to one of the authors of this article. The complete list of these URLs can be retrieved from the following URL: http://www.cs.concordia.ca/$\sim$faculty/bcdesai/search-oct97/whereis-Desai.html.

The first set of tests, the results of which are given in Table 2, was done on the following search engines: Alta Vista, Excite, Hotbot, Infoseek, Lycos, OpenText, and Yahoo.

Table 2.
SEARCH STATISTICS FOR USING THE SEARCH TERM BIPIN (AND) DESAI: SEPT. 1997

Search System	*Number of Hits*	*Number of Duplicates*	*Number of Mis-hits*	*Number of Defunct*	*Number of Items Missed*
AltaVista/ Yahoo	97	9	23	4	264
Excite	114	10	29	7	247
InfoSeek	8	2	1	1	319
Lycos	57	7	15	14	297
Hotbot	247	28	58	19	155
OpenText	19	-	7	5	318

As in the 1995 series of tests, we have shown the results by noting the number of hits produced, the number of duplicates, number of mis-hits, and the number of relevant documents not listed in the result; we have also included a column for the number of defunct URLs (which do not lead to any valid target Web pages). The duplicates are either the same document being served from two sites or the same document from the same site listed more than once. The latter errors seem to have been corrected in most search engines which do sufficient pre-processing of the result to eliminate obvious duplicates before presenting it to users.

The documents missed could be due to the approximations used by engines such as Alta Vista when it finds a large number of hits. However, the fact that these search engines could not locate all documents indicates the difficulty of reaching isolated URLs by search robots.

The bigger problem is the lack of selectivity and the measure of usefulness of the documents found by the search engines. We have collated the results by following the trail of "next" sets of URLs, and these could be viewed by pressing on the number of hits for each search engine in the online version of Table 2 (Desai, 1997a). A glance at the abstract or sum-

mary presented by the search engine indicates that they are not very revealing and, except for the most pedestrian needs, following the pointers would result in a drain of the searcher's time.

Search Statistics for Using Various Search Strategies

In a third series of tests, we used a simple search with the search terms: Bipin Desai, the advanced search expressions "Bipin Desai," and "Bipin C. Desai" respectively. These tests were made only on Alta Vista/ Yahoo. The results of these tests are given in Table 3.

Table 3.
Search Statistics for Using the Various Search Terms: Sept. 1997

Search System	*Number of Hits*	*Number of Duplicates*	*Number of Mis-hits*	*Number of Defunct*	*Number of Items Missed*
Alta Vista/ Yahoo	4285	30-90%	10-80%		200+
Alta Vista/ Yahoo	29	2	13	3	312
Alta Vista/ Yahoo	128	14	-	10	201

The result for a simple search of Bipin Desai (row 1 of Table 3) shows a high number of hits (4,285 in the test reported here; there is a bit of variation due to Alta Vista's method of abandoning a search after a sufficiently large number of hits is made). However, the simple search produces very low selectivity and relevance. Most of the hits in the top 160 entries are irrelevant, and a large number of relevant documents are not located. Most searchers will not have the patience to go through more than a few pages of the result, there being some 214 pages of the result for 4,285 hits.

The result for an advanced search expression for "Bipin Desai" (row 2 of Table 3) gives a lower number of hits and relevance since the author prefers to include his middle initial in the name. Most searchers may not be aware of such details.

The result for an advanced search expression for "Bipin C. Desai" (row 3 of Table 3) gives a relatively large number of relevant documents, some of which are duplicates, being accessible from more than one site. Some of the defunct URLs are not deleted by the search engines, pointing to the maintenance problem of the underlying database. However, this search still missed about two-thirds of the documents.

These tests lead us to believe that a search system should support better semantics. It is our opinion that the semantic header-based system (see Figure 2) (Desai, 1997b), wherein the provider of the resource is responsible for generating the entry, would be a more useful scheme to

support discovery. The semantic header is designed to describe the semantic contents of the source information resource and is better suited to supporting knowledge discovery than the actual resource. Many formats of a resource may not be directly accessible electronically, be suitable for direct discovery, or may require a considerable amount of computation and extremely slow response. The semantic header could also be used as a surrogate to express semantic dependencies inherent in a collection, which is not possible to do with existing search engines.

The quality and the reliability of the document could be expressed by including reviewers' comments in the form of annotation with the semantic header. Such reviews are rarely accessible in traditional cataloging systems. However, in the CINDI system this, along with the abstract supplied by the authors, would be valuable in judging the suitability of a document to a searcher. It could also give feedback to the provider. The semantic header metadata also allow the server system to perform initial query processing and thus reduce the cost involved in accessing and processing irrelevant resources.

Overview of the CINDI System

The overall structure of the CINDI system is shown in Figure 1. The workstation at the provider's site contains the CINDI client software and a partial catalog. The client software is composed of a registering graphical user interface, the client portion of a distributed expert system, and the associated knowledge base. The semantic header information entered by the provider of a resource using this graphical interface is relayed from the user's workstation by a client process to the database server process at one of the nodes of the SH Distributed Database (SHDDB). The node is chosen based on its proximity to the workstation or on the subject of the index record. On receipt of the information, the server verifies the correctness and authenticity of the information and, on finding everything in order, sends an acknowledgment to the client. It also has a partial catalog of the thesaurus database. The function of these are described later in the section on the Semantic Header Registration System.

The server node is responsible for locating the partitions of the thesaurus for the subject hierarchy or the sites of the SHDDBs where the entry should be stored and forwards the replicated information to appropriate nodes. The server node is also responsible for providing the catalog information for the search system. In this way, the various sites of the database work in cooperation to maintain consistency of the replicated database. The replicated nature of the database also ensures distribution of load and continued access to the system when some sites are temporarily nonfunctional.

The user interface for the CINDI system consists of three graphical interfaces: the SH index registration system, the search system, and the

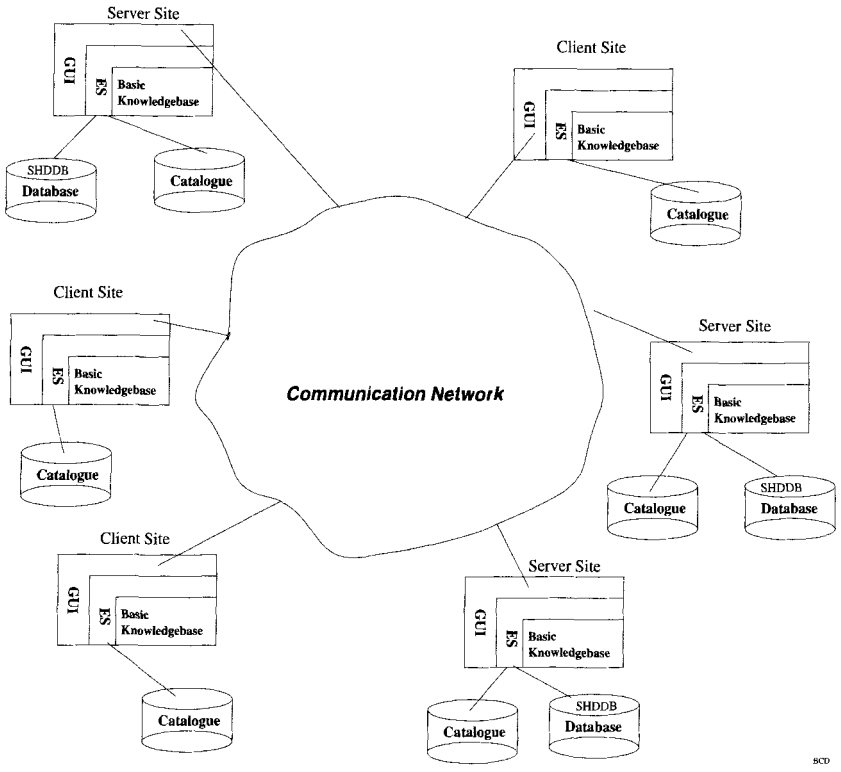


Figure 1. Overall Structure of the CINDI System.

annotation systems. The SH index registration allows a document provider to fill in the slots of the SH. The search interface is used by a searcher to locate documents. The annotation interface allows a user to insert comments on a document in its semantic header. The indexing and search systems have an associated expert system that helps the providers and searchers in selecting appropriate subject terms to best describe the source document or the query respectively. These functions are described briefly later in this article.

The SHDDB contains information on subject hierarchy, a thesaurus to help select controlled terminology from the subject hierarchy, the registered semantic headers, and the associated annotations. An expert system mimics the cataloging librarian and helps a provider make suitable controlled subject entries in the semantic header. When a searcher uses the expert system, it mimics a reference librarian in helping him locate the relevant subjects. The communication from a user workstation to the database site uses the client/server paradigm.

The slots of the SH, as seen in Figure 2, contain the title of the document, its authors, the subject(s), abstract, and so on. The intent of the semantic header is to include those elements that are most often used in the search for an information resource. Furthermore, the SH provides information on the organization of a document such as chapters, sections, or whether the document is part of or an actual collection. The registry containing all the semantic headers is much smaller than the actual collection of documents (Desai, 1997). A person searching for a document

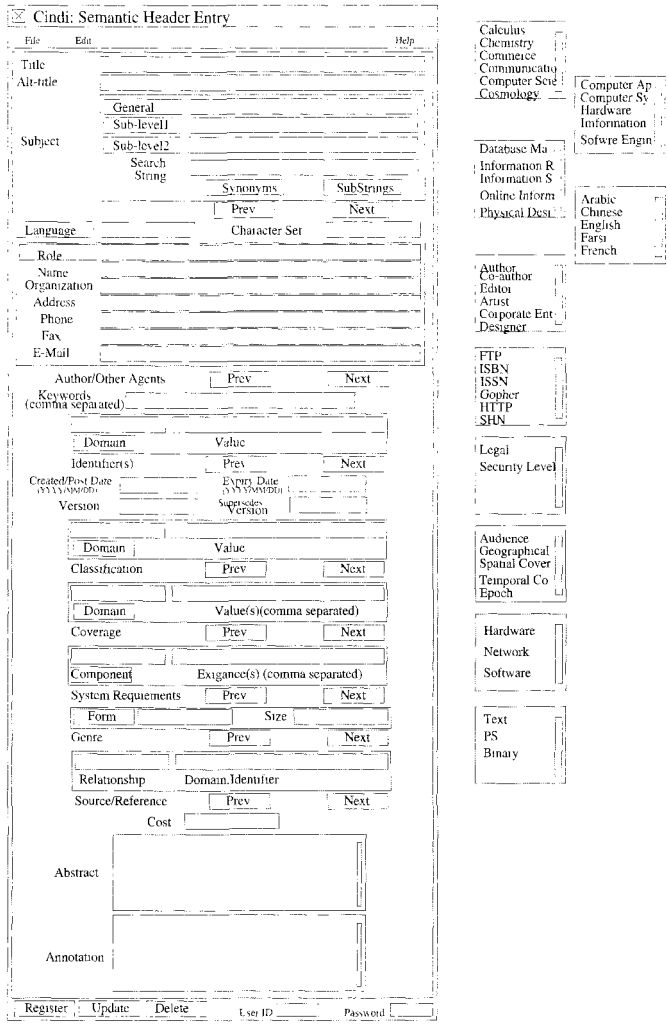


Figure 2. Components of the Semantic Header.

first locates the appropriate SHs. Once these are found, the actual documents can be easily accessed. Since the registry for the SH is smaller than the actual collection, and much of the query for the required search terms could be preprocessed, searching becomes faster.

Semantic Header Registration System

The Semantic Header entry and registration subsystem provide a graphical interface (similar to Figure 2) to facilitate the task of the provider (author/creator) of a resource to register the SH for the resource. The system also offers help by means of pop-up selection windows and an expert engine to suggest controlled terms. This expert engine is intended to bring some of the expertise of a catalog librarian to the ordinary user (Chander, Shinghal, Desai, & Radhakrishnan, 1997). Many of the elements in the SH can be extracted directly from the resource document if they are properly tagged (the Automatic Generation of Semantic Header project is currently underway at Concordia). Once the information is correctly entered, the provider can decide to register the SH entry in the database. When the SH information is accepted by the database, the provider is notified. A user ID and the associated password is required when the SH is first registered and for all changes made to it. Since the user ID and the password are not accessible by anyone other than the original registrar (usually the provider) of the index entry, the entry can only be updated by persons who are cognizant of them. Changes that may be made could be due to changes made in the resource or its migration from one system to another. A copy of the SH is stored at the provider's site for convenience in later updates.

The subject for the document being indexed is selected hierarchically. The provider first selects the general level, and a rule-based system thereafter guides the selection of the corresponding lower levels. In case the provider is unclear about the subject area of the document, she can seek help by entering a string in the "search string" slot and use either the synonym or substring push button. A rule-based system is invoked and guides the provider in selecting the appropriate subject. The provider is not allowed to enter the subject terms directly, thus restricting the subject terms to a controlled vocabulary from the subject headings.

Some of the slots can contain more than one value—e.g., the author slot can have more than one name and address to signify that the document has multiple authors. These multiple values are entered by using the NEXT and PREVIOUS buttons in the corresponding slots. Using this scheme, any number of values for such slots can be entered.

When a new provider first creates an SH, she chooses, with the help of CINDI, her own unique user ID and password. These values are stored

in the CINDI database and are associated with all SHs registered using this pair of values. The values of the SH slots can be updated only by the provider of the SH. The only exception is the annotation slot: here, any user can insert comments. Only the provider is allowed to delete her semantic header. The deletion of an SH would not affect its user ID and password.

Expert System Support for Registering

Expert and knowledge based systems have been used in various domains to provide users with the expertise of a domain expert. In our system, we need to guide the provider of a document in choosing the appropriate subject hierarchy from a controlled hierarchy (we use a hierarchy derived from ACM, INSPEC, and the Library of Congress Subject Headings [LCSH]). In many domains, the domain knowledge is encoded as a set of "if . . . then . . .else" rules. Encoding knowledge in such a manner and checking the user input against such encoded rules has been found to be fairly inefficient. Our initial approach using CLIPS (Giarratano & Riley, 1994) to encode these rules proved this observation. Furthermore, CLIPS imposed a considerable overhead on the system.

In registering an SH, and for later searching, it is essential to employ the knowledge and expertise of cataloging librarians. However, employing professional librarians may be costly, thus the need of an expert system to model librarians' expertise and guide users in cataloging and searching. The expert system would help users choose correct subject terms. It would also guide them to register, update, delete, and annotate SHs. The expert system is designed so that its query for resource searches facilitates efficient database access and reduces the number of incorrect results generated.

Expert systems have been used to encode the expertise of experts in well-defined domains. The system can then be used to guide users in reaching the same conclusion as the domain experts in a given situation. In our case, we need the expertise of a cataloging librarian to help users choose appropriate controlled terms using a knowledge-base encoded as a thesaurus of synonyms.

A typical user wanting to create a semantic header entry for her document usually does not have precise knowledge of the exact subject heading hierarchy under which the document should be classified. However, she has a very good idea of the exact topic(s) treated in her document and knows the usual terms used in the relevant literature. Such terms may not be the same as the controlled terms established by a cataloging authority such as the LCSH. The cataloging system, mimicking a cataloging librarian, should be able to guide the user in a search for controlled terms from a subject heading hierarchy.

Checking for all input combinations using direct encoding of knowledge has been found to be very inefficient (Chander, 1995). Our first attempt was to incorporate the expert system rules by embedding an expert system shell in the cataloging system. However, this approach was not only slow but increased the overall size of the system. Furthermore, the shell was not sensitive to the context under which a rule was to be tested and hence had to recompute the set of rules to be tested.

Our second attempt was to replace the expert system shell by distributing the rules in the appropriate components of the system. This not only increased the context sensitivity of the system and reduced the size of the program code, but it also reduced the number of rules that had to be considered for each subsystem. This distributed system was encoded directly as C/C++ functions, thus further improving the system performance. In this prototype version of the system, we allowed the user to enter a term at any of the three levels of the subject hierarchy. If the term entered was found to be a synonym of a controlled term at the same level as the one entered by the user, then the system would show the controlled term and the corresponding higher level term(s) and prompt the user to confirm them. However, should the term entered by the user be a synonym of a controlled term at a different level of some subject hierarchy, then the system requests the user to resolve this conflict and warns the user of this inconsistency. Also, a term entered by the user at a lower level may not be a synonym for any term at this level of the current subject hierarchy. To avoid such confusion and the possible need to backtrack, we revised our scheme to take advantage of the interactive and graphical nature of the interface.

In our final implementation, our strategy is to separate the subject hierarchy search into two orthogonal components: (1) a strictly hierarchical

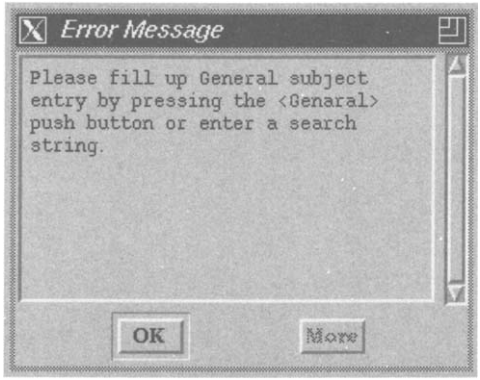


Figure 3. Selecting Subject Level-1 Before the General Level.

subject entry system using context sensitive pull down menus, and (2) finding controlled terms for a string or term entered by a user. In the first case, the user is guided by the system to first select a higher level subject before being allowed to select a lower level. In this component, the user is not allowed to directly enter a subject term as indicated in Figure 3. In this way, the search by the system for a lower-level is thus limited to the already selected controlled higher levels and avoids the confusion and backtracking.

In the event a resource provider has a basic idea of a term in the subject hierarchy that she or he wants to use as a resource, the term could be entered in the search slot of the GUI and use the synonym substring feature to query CINDI for a controlled term. Terms entered in this manner by a user could be synonymous at any level of a subject hierarchy. This could also occur when a provider enters terms that are part of controlled terms (substrings) and thus cause the term to match entries at multiple levels. If the provider enters a synonym such as *system*, which could occur at a large number of subject hierarchies, the expert system displays the result and suggests the provider enter a more specific synonym. The system, using a thesaurus of synonyms and controlled terms, finds subject hierarchies closest to a user's entered term and presents these for selection. The displayed term could be at any level of the hierarchy. The result of the search is presented to the provider in a pop-up window, as shown in Figure 4; the provider would then make a selection at one of the levels indicated by the system or reject the choice to try another term. The provider views the matching subject terms for each matching level by selecting the corresponding push button in Figure 4. If the entered term maps into a controlled term at a lower level of a subject hierarchy, the system automatically selects the higher levels of the subject hierarchy and displays them for the user to make a selection. These are illustrated in Figures 5 and 6.

The use of the GUI, along with the orthogonal separation of subject level entry in hierarchical order and by matching a synonym, has simpli-

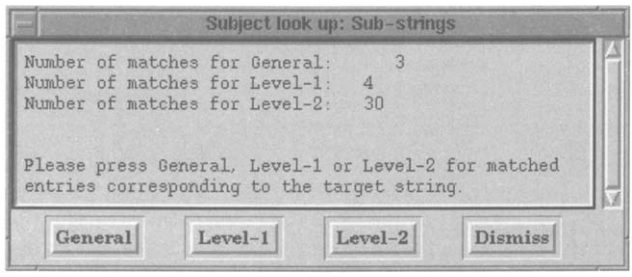


Figure 4. Example of Sub-String Look-Up: Search String "Com."

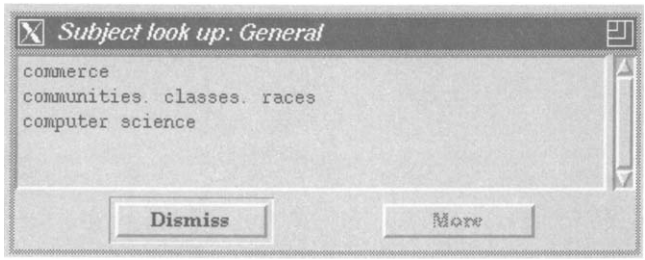


Figure 5. Example of Sub-String Look-Up: Display General Level.

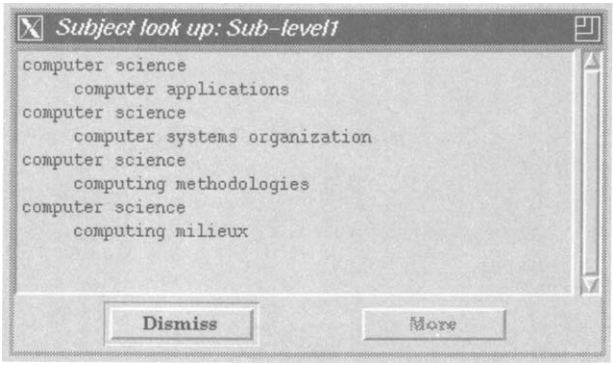


Figure 6. Example of Sub-String Look-Up: Display Sub-Level-1.

fied the implementation of both the cataloging system and the searching system. By limiting the provider to enter a higher level subject before a lower level subject (see Figure 3), we have avoided many pitfalls that occurred in our earlier implementation and hence simplified the set of rules and the data that have to be retrieved. This improved the performance of the system and allowed us to provide context-sensitive help for indexing, updating, and searching.

Some fields of the semantic header may have multiple or repeatable entry fields. The repeatable entry fields in the SH allow, for example, a document to be classified under more than one subject; an article could be written by more than one author; and the document can be identified by its HTTP, FTP, ISBN, and so on. We provide PREV and NEXT push buttons (Figure 2) at the bottom of each block in the user interface to accommodate these entries.

The PREV push button allows the user to view or modify the previous entry of a block and the NEXT push button allows the user to enter, view, or modify the next entry of a block. To proceed to the previous or next entry, certain rules are enforced. For example, the NEXT entry button

action is not allowed until the current entry is completed and it has a valid value. The validity of all slot values is verified at the client site before the system allows the semantic header entry to be registered.

CINDI Database System

The index entries registered by a provider of a resource in CINDI are stored in SHDDB. From the point of view of the users of the system, the underlying database may be considered to be a monolithic system. In reality, it could be distributed and replicated, allowing for reliable and failure-tolerant operations. The interface hides the distributed and replicated nature of the database. The distribution is based on subject areas and, as such, the database is considered to be horizontally partitioned (Desai, 1990).

The database on different subjects in the CINDI system is to be maintained at different sites of the communication network (as illustrated in Figure 1). The locations of such nodes need only be known by the intrinsic interface and the database catalog used to distribute this information. Catalogs would also be used to store information about the location of the subject areas maintained in the SHDDB so that the client process at the user workstation can select subject hierarchies for indexing and retrieving semantic headers.

The semantic header entered by the provider of the resource using a graphical interface is relayed from the user's workstation by a client process to the database server process at one of the nodes of the SHDDB. The node is chosen based on its proximity to the workstation or on the subject of the index record. On receipt of the information, the server verifies the correctness and authenticity of the information and, on finding everything in order, sends an acknowledgment to the provider at the client site.

The thesaurus database contains four object classes which represent the general subject of the subject hierarchy, the sub-subject and the sub-sub-subject, and finally the *synonym* which contains those subject terms (at any level) synonymous with the controlled terms. The registration subsystem at the server site is responsible for registering, deleting, updating, and annotating semantic headers. This subsystem uses an SHDDB area to store the SHs and other related objects. The database is an aggregation of three objects: SH, user ID, and word. SH contains all fields of the semantic header. Some of these fields are included in the SH object as attributes; others are objects which are components of the SH object. As an example, the *author* object is a part of the SH object which should be (partially) ordered. This is because, in the GUI, the first author field entered by the user would take part in the construction of the semantic header name (SHN). The SHN is derived from the following required elements in the semantic header: title, name of first author (or name of

organization, if the document or resource being registered is attributable to a corporate or organizational entity), first subject, creation date, and version. The SH also includes information in the Identifier object to access the corresponding online information resource. The UserID object contains both a user ID and a password entered by the user in the GUI.

The SHDDB also contains the *word* object. It stores non-noise words appearing in those fields of an SH which may be used during the search operation. The *word* object corresponds to the SH objects where the value of the *word* object appears in the semantic headers. This object contains a fixed number of *context* objects equal to the number of search GUI fields.

The semantic header database and the catalog database are implemented using the ODE (Object Database Environment) database system (Arlein, Gava, Gehani, & Lieuwen, 1993; Agrawal & Gehani, 1989; Agrawal, Dar, & Gehani 1993; Biliris & Panagos, 1993).

Registering the Semantic Header

The graphical interface (see Figure 2) facilitates the provider (author/creator) of a resource to fill in and subsequently register the bibliographic information about the resource. Once the information is entered, the provider can decide to register the semantic header entry in the database. The REGISTER push button allows providers to register the current semantic header into the CINDI database. To register a new semantic header, a user ID and password are required. Before an actual registration request is made to the server, the client system would check the SH entry to ensure that all the required fields are entered. This validation ensures that the standard indexing scheme is enforced.

When a semantic header is received at the database (server) site, the system performs a number of operations to register an SH. The first step is for a parser to verify the syntax of the input file and ensure that the mandatory fields of the SH have been entered. The non-noise words of the semantic header are stored in temporary variables and data structures for later use. The next step is for the database module to verify the status of the user ID and the password and ascertain that the SH does not already exist in the database. Finally, the words and the semantic header are indexed into the database. The non-noise words would be added to the database and all attributes of the SH object would be initialized. Finally, the unique SHN identifier is assigned to the newly added SH object. In a case where an error occurs, an error code would be sent to the client site which would be used by the client expert system to guide the user to correct the problem.

When an SH is registered by the server, a copy of it is stored locally at the client site. Later, this could be loaded to update the semantic header as discussed in the section on Updating a Registered Semantic Header.

Client-Server Communication

The communication between the user interface and the database is made using the TCP/IP protocol written in C. The server daemon runs at the server site. When a client, at the user site, is called by the user interface, it connects to the server and sends data in a file containing the query request. The server calls appropriate functions for parsing the file and transforming it into a database specific query (or queries). This query is sent to the database for processing. Finally, the server receives the result of the query in a file created by the database module and sends it back to the client using the TCP/IP protocol.

Since a server may provide services to more than one client at a time, the server assigns a unique client ID to each file received from the client site. Each client ID is a concatenation of three fields. The first field is a fixed string used in all files. The second field is the value of time in seconds. The third one is the process ID of the child process responsible for serving a specified client. Thus possible client ID collisions at the server site are avoided.

In a case where a network problem prevents data transmission, the server program provides a timeout mechanism to prevent the GUI at the user site from waiting for the server to respond indefinitely. If, after a specified period of time, the server fails to complete the process of transmitting data from/to the client, the server sends an appropriate message to the client process and disconnects from the client process. Subsequently, the user interface receives the error code from the client process and displays an error message to the user. This way, the GUI does not freeze, and the user can carry on making other requests.

Once a client has established a connection to a server, it issues a series of transactions each of which is invoked by a function call. One such transaction actually performs the connection between client and server. With the exception of the connect transaction, each transaction generates a sequence of actions as follows: a message identifying the transaction to be executed and the associate data are sent by the client to the server; the server process identifies the transaction from the associate data; and a message containing the response to the transaction is sent by the server to the client along with the ID of the original transaction. This is repeated until the client initiates a terminating transaction.

Updating a Registered Semantic Header

The UPDATE push button (see Figure 2) allows a provider of resources to update an existing semantic header. When there is a need to update the SH, it is loaded from a local saved copy into the GUI for registration through the use of the OPEN menu item in the FILE menu. The provider is not allowed to change fields corresponding to the SHN nor the reviewers' annotations (except for the annotations made at the time of

registration; the others are not stored locally in the original copy of the registered SH at the client site). The GUI rule system enforces this requirement by making these fields non-editable during the update phase and by giving appropriate warning messages if the provider tries to modify these fields. To register the updated semantic header, the provider has to enter the user ID and password that were used when initially registering it.

The UPDATE button is disabled when a new semantic header is being entered. The REGISTER push button is disabled when a provider opens an existing semantic header for modification.

Deleting a Registered Semantic Header

The DELETE push button is used to remove a semantic header from the system. An SH can be deleted by the provider who registered it only if no public annotation were made in it after the SH was registered. The procedure of deleting an SH is similar to that of updating it. To delete an SH, the provider enters both a user ID and the corresponding password that match those entered when the SH was registered. Otherwise, an error message is relayed to the provider. The SH and the SHN maintained in the word object with the value corresponding to each non-noise word are deleted from the database. If a word object, after such deletion, is found to be associated with no SHNs, it would be deleted from the database as well. The DELETE button is disabled when the user enters a new SH.

SEARCHING

In the current search system, we have incorporated the elementary expertise used by a reference librarian. Reference librarians are aware of the conventions used by cataloging librarians. They are conversant with the classification schemes, terms, indexes, structures, and resources available for a user's particular need. This basic expertise of the librarian is replicated to assist the users of our application in discovery and guides the user in entering the various search items in a graphical user interface similar to the one used by the registering subsystem (see Figure 7). The system is designed so that its query for document search facilitates efficient database access and reduces the number of incorrect results generated. For example, the system aids the user in completing a given field entry based on the contents of the other search fields. As in the case of the registering sub-system, the expert system provides context-sensitive help in choosing appropriate search terms for index entries such as the subject, sub-subject, sub-sub-subject, and so on (Chander et al., 1997). When no SHs are found, the system suggests other alternatives to the user.

The CINDI system offers a wide range of search criteria to allow precisely targeted resource retrieval. Most widely used search fields are

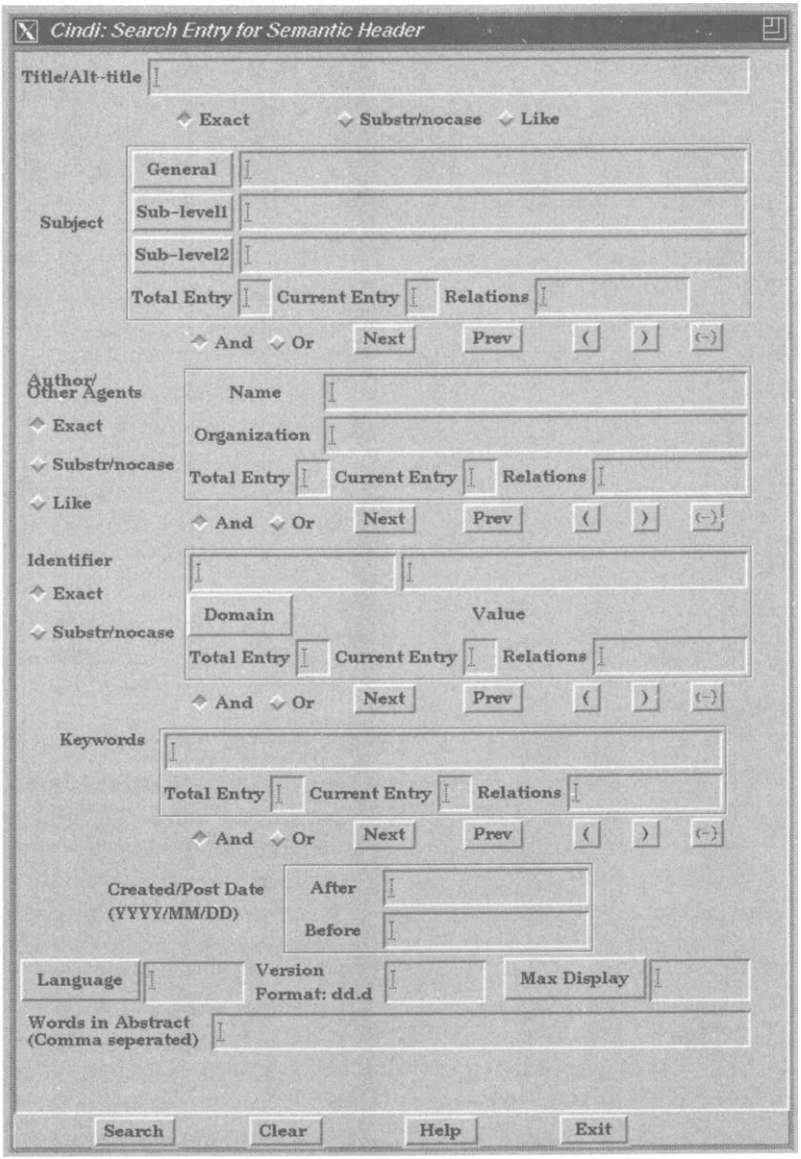


Figure 7. Graphical User Interface: Search System.

included to allow the user to tailor the search and discovery needs. In most of these fields, the search can be specified to be performed using an exact match or substring of the word entered by the user. Most of these fields allow the logical operations *and* and *or* to refine the search. Parentheses are also provided to allow nested logical search predicates. To transform the user-defined queries received from the client into database queries, we employ the reverse Polish (postfix) notation.

Once the user has entered a search request, the client process communicates with the nearest server which determines the appropriate sites of the SHDDB. Subsequently, the server communicates with these sites and retrieves one or more SHs. The results of the query can then be collected and a list in user specified block size is sent to the user's workstation as illustrated in Figure 8. The contents of any of these semantic headers are displayed on demand by clicking on the title in the list. This is illustrated in Figure 9. The user can navigate to the other semantic headers by pressing the appropriate push buttons in this display. The system allows the user the facility to access one or more of the actual resources by connecting to a browser. A user can access the actual resource only if the selected item for access has an identifier which allows online access via a browser such as Netscape which could be used to display the resource.

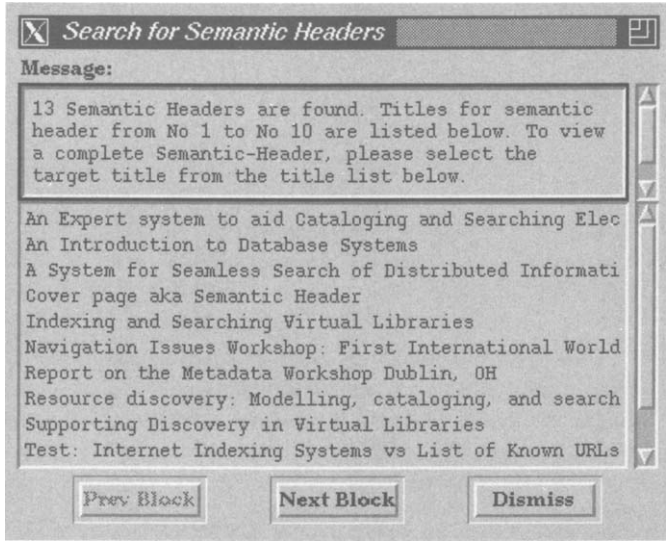


Figure 8. Result of a Search Query: List of Semantic Headers.

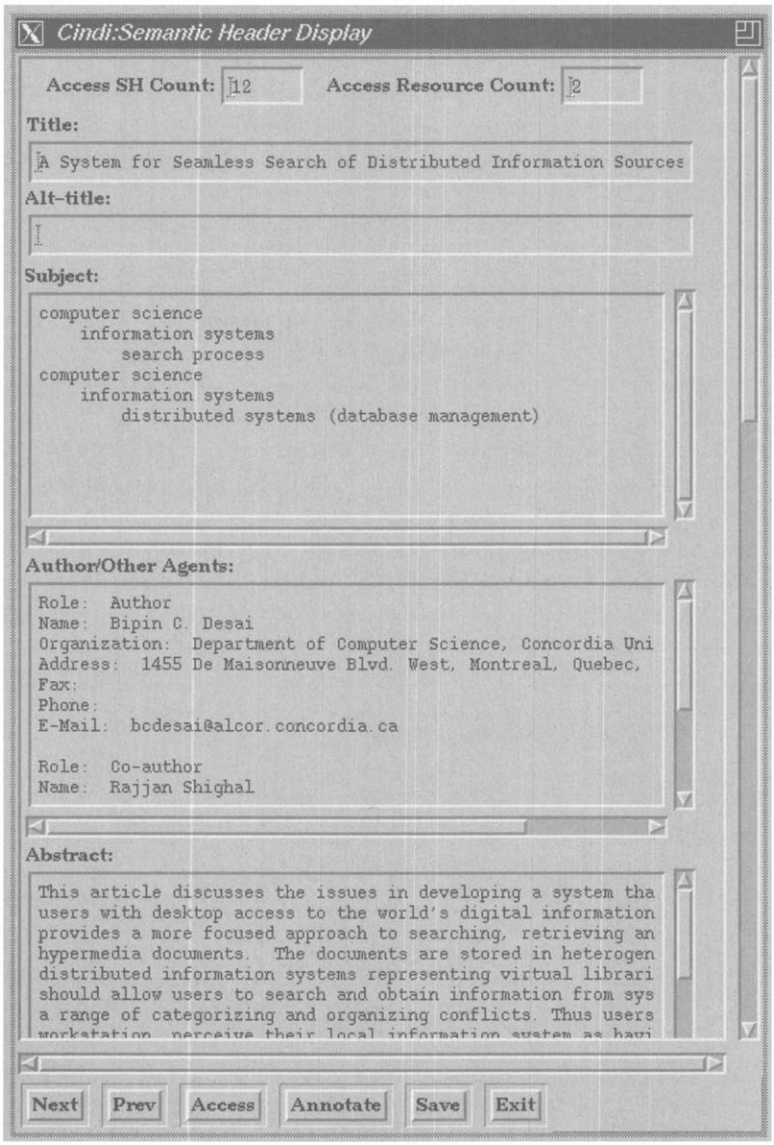


Figure 9. Display of a Selected Semantic Header.

We associate two counters with each SH to measure the extent of the dissemination of the SH and the resource it describes. One counter records the number of times a given SH was accessed in searches made by users of the system. The second counter indicates the number of accesses made via CINDI to the actual resource corresponding to a semantic header.

ANNOTATION

The research community depends on peer review of documents submitted for publication. Such review or annotation is often not published. However, comments to the editor made by readers of journals are usually published and are accessible to the community. Since many of the resources on the Internet tend not to be reviewed, it would be beneficial for a user to have access to annotations made by other users for a given resource. The proposed system allows users to add annotations to an existing resource.

These annotations are stored along with the index in the SHDDB (Desai, 1996). Peer reviews of electronically submitted papers could be implemented using such annotations. Authentication of reviews has to be done by an appropriate editorial board.

The graphical user interface of the annotation subsystem is shown in Figure 10. The annotation subsystem is similar to the indexing subsystem. However, only a few of the indexing entries that uniquely identify the resource in question are required. An annotation made by any user can be entered and would be registered with the identity of the user. Such annotations could be valuable guides for future users.

To avoid nonserious entries to the annotation by unscrupulous readers, we have separated the annotation entry from the search subsystem. In order to add annotations, the user has to save the semantic header when viewing it in the search subsystem using the SAVE push button. It is expected that the user would actually access the resource corresponding to the SH. If the user decides to make an annotation, he loads the saved SH from the local file system by pressing on the LOAD SH push button in the annotation form.

The newly entered annotations, together with the annotator's information, as well as his login name and host name, would be concatenated to the existing annotations when the annotator registers them.

CONCLUSION

Current index systems are based on harvesting the network for new documents. Such documents are retrieved and their contents used to provide terms for the index. The big disadvantage with this scheme is the unreliability of the index entries produced and the lack of an authentic abstract for the item. The current Dublin Metadata Element (Desai, 1995) list also suffers from the absence of the abstract. Furthermore, current index schemes are relevant for resources of limited protocol and are not applicable to other resources. CINDI has addressed these problems, giving rise to the following advantages: CINDI allows the indexing of resources accessible online or offline; CINDI requires that the provider of the resource use controlled terms and provide an abstract (this is an improvement over extracting phrases from the first part of a resource or by simply

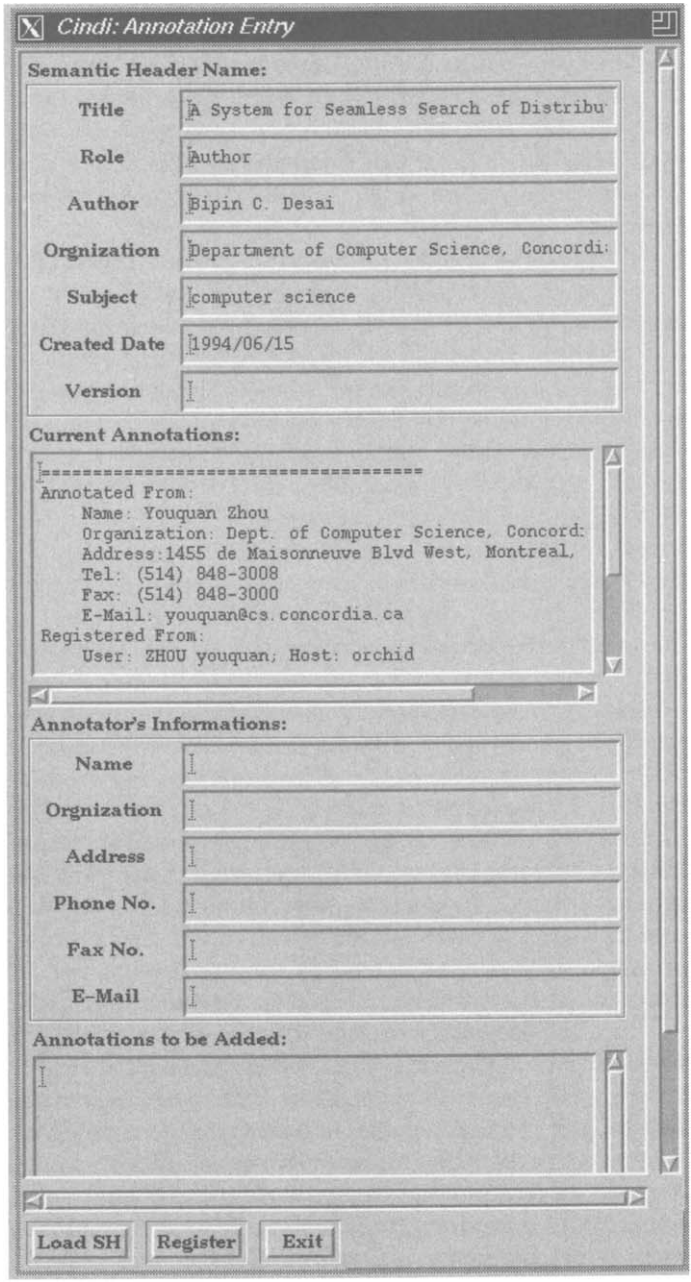


Figure 10. Graphical User Interface: Annotation System.

picking up terms by scanning a resource as is done by some of the existing systems), since the registration of the semantic header in the database is performed by the provider of the resource, it improves cost, accuracy, and efficiency; CINDI allows annotations by reviewers which enable future users to make better informed decisions regarding the relevance of the source resource; the size of the CINDI database is not limited since the database is distributed among a number of sites. Furthermore, the system lends itself to the well researched distributed query processing techniques to support discovery from these distributed database systems in parallel.

The expert system support provided in our implementation is obtained by distributing the rules to be enforced in various parts of the system including the GUI. This method was chosen over using expert system shells such as CLIPS. While such shells facilitate knowledge engineering and rule encoding, we found in our trial implementation that they incurred a significant overhead. For example, for every rule firing, if a system such as CLIPS were used, its inference engine would recompute the set of rules that can fire. The distribution of the rules in CINDI improved performance since only a small number of rules had to be developed in such an environment.

CINDI, in its current implementation, uses an ODE database and an X-window based Motif interface (Heller, 1994). The client software for ULTRIX for Sun can be downloaded from: <http://cindi.cs.concordia.ca/cindi>. We plan to port the client software to Linux. The initial decision to use a Motif-based interface was due to the limitation of HTML (Hypertext, 1997) in providing unspecified numbers of repeating fields. With the newer tools, such as XML and Java, we have undertaken porting CINDI to the Web. We will also port SHDDB to a robust commercial DBMS.

CINDI, as do other self-indexing systems, requires the active participation of the provider. To make this task easier, we are providing an automatic semantic header generation system. Preliminary results of this work in progress is encouraging, and it will be incorporated in our Web-based version.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of P. G. Chandra for developing the initial expert systems for cataloging librarians. Concordia University reference librarians Carol Coughlin and Lee Harris were always available when we had questions about cataloging and reference librarian practice. This contributed greatly to the rules developed for the expert system and the thesaurus. We are also grateful to the editors, Jian Qin and M. Jay Norton, for their very constructive comments which helped improve the paper. In addition, this work is partially supported by grants from NSERC.

REFERENCES

Agrawal, R., & Gehani, N. (1989). ODE (object database and environment): The language and the data model. In *Proceedings of the 1989 ACM-SIGMOD International Conference on the Management of Data* (Portland, Oregon, 2 June 1989) (pp. 36-45). New York: Association for Computing Machinery.

Agrawal, R.; Dar, S.; & Gehani, N. (1993). The O++ Database programming language: Implementation and Experience. In *Proceedings of the Ninth International Conference on Data Engineering* (April 19-23, 1993, Vienna, Austria) (pp. 61-70). Los Alamitos, CA: IEEE Computer Society Press.

Arlein, R.; Gava J.; Gehani, N.; & Lieuwen, D. (1993). *Ode 4.1* (user manual). AT&T Bell Laboratories.

Biliris, A., & Panagos, E. (1993). *EOS user's guide* (Release 2.0). AT&T Bell Laboratories.

Brody, H. (1995). Internet@crossroads.$$$. *Technology Review, 98*(May/June), 24-31. Retrieved from the World Wide Web January 11, 1999: http://www.techreview.com/articles/may95/Brody.html.

Brownlee, N. (1995). *New Zealand experiences with network traffic charging.* Retrieved January 11, 1999 from the World Wide Web: http://www.auckland.ac.nz/net/Accounting/nze.html.

Byrne, D. J. (1991). *MARC manual: Understanding and using MARC records.* Englewood, CO: Libraries Unlimited.

Chander, P. G.; Shinghal, R.; & Radhakrishnan, T. (1995). Goal supported knowledge base restructuring for verification of rule bases. In R. F. Gamble (Ed.), *IJCAI'95 Workshop on Verification and Validation of Knowledge-Based Systems, Montreal* (pp. 15-21). Somerset, NJ: IJCAI, Inc.

Chander, P. G.; Shinghal, R.; Desai, B. C.; & Radhakrishnan T. (1997). An expert system to aid cataloging and searching electronic documents on digital libraries. *Expert Systems with Applications, 12*(4), 405-416.

Cocchi, R.; Estrin, D.; Shenker, S.; & Zhang, L. (1991). A study of priority pricing in multiple service class networks. Retrieved January 11, 1999: ftp://parcftp.xerox.com/pub/net-reserch/pricing1.ps.Z

Crawford, W. (1984). *MARC for library use: Understanding USMARC formats.* Boston, MA: G. K. Hall.

Cromwell, W. (1994). The core record: A new bibliographic standard. *Library Resources & Technical Services, 38*(4), 415-424.

Desai, B. C. (1990). *Introduction to database systems* St. Paul, MN: West.

Desai, B. C. (1994a). *The semantic header and indexing and searching on the Internet.* Retrieved January 11, 1999 from the World Wide Web: http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html.

Desai, B. C. (1995a). *Test: Internet indexing systems vs list of known URLs.* Retrieved January 11, 1999 from the World Wide Web: http://www.cs.concordia.cs/~faculty/bcdesai/test-of-index-systems.html.

Desai, B. C. (1995b). *Report of the Metadata Workshop, Dublin, OH (March 1995).* Retrieved January 11, 1999 from the World Wide Web: http://www.cs.concordia.ca/~faculty/bcdesai/metadata/metadata-workshop-report.html.

Desai, B. C. (1997a). *Test: Internet indexing systems vs list of known URLs: Revisited.* Retrieved January 11, 1999 from the World Wide Web: http://www.cs.concordia.ca/~faculty/bcdesai/revisited.html.

Desai, B. C. (1997b). Supporting discovery in virtual libraries. *Journal of the American Society of Information Science, 48*(3), 190-204.

Desai, B. C., & Shinghal, R. (1996). Resource discovery: Modelling, cataloging, and searching. In *DEXA '96* (Seventh International Workshop on Database and Expert Systems Applications, September 9-10, 1996, Zurich, Switzerland) (pp. 70-75). Los Alamitos, CA: IEEE Press, Zurich, Switzerland.

Gaynor, E. (1994). Cataloguing electronic texts: The University of Virginia library experience. *Library Resources & Technical Services, 38*(4), 403-413.

Giarratano, J., & Riley, G. (1994). *Expert systems: Principles and programming* (2d ed.). Boston, MA: PWS Publishing.

Giordano, R. (1994). The documentation of electronic texts using Text Encoding Initiative headers: An introduction. *Library Resources & Technical Services, 38*(4), 389-401.

Heller, D., & Ferguson, P. M. (1994). *Motif reference manual.* Sebastopol, CA: O'Reilly & Associates.

Horny, K. L. (1985). Minimal-level cataloguing: A look at the issues. *Journal of Academic Librarianship, 11*(6), 332-334.

HyperText Markup Language homepage. (1997). Retrieved January 11, 1999 from the World Wide Web: http://www.w3.org/pub/WWW/MarkUp/.

Kahle, B. (1991). *An information system for corporate users: Wide area information servers* (Thinking Machines Technical Rep. No. TMC-199).

MacKie-Mason, J., & Varian, H. (1997). *Usage-based pricing: Analyses of various pricing mechanisms.* Retrieved September 1997 from the World Wide Web: http://gopher.econ.lsa.umich.edu/EconInternet/Pricing.html.

Mauldin, M. L. (1995). *Measuring the Web with Lycos.* In B. C. Desai & B. Pinkerton (Eds.), *Proceedings of the WWWII Workshop on Web-wide indexing/semantic header or cover page* (Darmstadt, Germany, April 1995) (pp. 26-29). Retrieved August 4, 1999 from the World Wide Web: http://www.cs.concordia.ca/~faculty/bcdesai/www3-wrkA/workshop-a.html

Petersen, T., & Molholt, P. (Eds.). (1990). *Beyond the book: Extending MARC for subject access.* Boston, MA: G. K. Hall.

Qin, J. (1998). *Computational representation of Web objects in an interdisciplinary digital library: A survey and experiment in polymer science.* Retrieved February 11, 1999 from the World Wide Web: http://www-dept.usm.edu//~slis/qin/metadata.

Rhee, S. (1985). Minimal-level cataloging: Is it the best local solution to a national problem? *Journal of Academic Librarianship, 11*(6), 336-337.

Ross, R. M., & West, L. (1985). MLC: A contrary viewpoint. *Journal of Academic Librarianship, 11*(6), 334-336.

Saunders, L. M. (1993). *The virtual library: Visions and realities.* Westport, CT: Meckler.

The Web robots pages. (1996). Retrieved January 11, 1999 from the World Wide Web: http://info.webcrawler.com/mak/projects/robots/robots.html.

Welcome to ALIWEB. (1995). Retrieved January 11, 1999 from the World Wide Web: http://www.nexor.co.uk/public/aliweb/aliweb.html.

Additional References

Desai, B. C., & Shinghal, R. (1994b). *A system for seamless search of distributed information sources.* Retrieved January 11, 1999 from the World Wide Web: http://www.cs.concordia.ca/old/w3-paper.html.

TEI guidelines for electronic text encoding and interchange. (n.d.). Retrieved September 1997 from the World Wide Web: http://etext.virginia.edu/bin/tei-tocs.

Abstracts and Abstracting in Knowledge Discovery

MARIA PINTO AND F. W. LANCASTER

ABSTRACT

VARIOUS LEVELS OF CRITERIA FOR JUDGING the quality of abstracts and abstracting are presented. Requirements for abstracts to be read by humans are compared with requirements for those to be searched by computer. It is concluded that the wide availability of complete text in electronic form does not reduce the value of abstracts for information retrieval activities even in such more sophisticated applications as knowledge discovery.

INTRODUCTION

Abstracts were first developed to be read by humans, providing concise summaries or descriptions of published items suitable for inclusion in printed indexing services or in scholarly journals along with the articles to which they relate. When computers started to have a serious impact on information retrieval in the 1960s, abstracts became important as human-readable output from electronic databases. Later, as storage and processing costs declined, they began to assume a new role—that of computer-searchable surrogates for larger bodies of text.

Today, of course, it is economically feasible to store vast quantities of text in computer-searchable form. Nevertheless, this has not made abstracts redundant. They remain useful summaries to be read by humans. Furthermore, if recall and precision are both taken into account, they may still be optimum for retrieval purposes because the searching of full text will frequently cause an unacceptable level of irrelevancy. Several

Maria Pinto, Departamento de Biblioteconomia y Documentación, Universidad de Granada, 18071 Granada, Spain
F. W. Lancaster, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, Champaign IL 61820
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 234-248

investigators (e.g., Tenopir, 1985) have shown that searching abstracts may be more effective or more cost-effective than searching of full text, while Salton (1971) found that, while full text gave better overall results than abstracts in the type of automatic processing employed in his SMART retrieval system, the differences were not great and the abstracts allowed more cost-effective processing.

On the surface, one might assume that knowledge discovery operations would be most likely to succeed when the complete text of items is processed. This is not necessarily so because full text can generate so many spurious relationships that significant and useful associations will be virtually impossible to recognize. Abstracts may still have great value in knowledge discovery activities as they do in many others.

This article will review various criteria by which the quality of abstracts may be judged. It will then discuss which criteria apply most clearly to the value of abstracts in knowledge discovery applications.

Quality in General

The word "quality" occurs frequently in everyday life and, in this general setting, stands for an idea that, while not necessarily exact, seems readily understood. On the other hand, in more formal and restricted applications—such as science, technology, commerce, and education—much less agreement exists on what "quality" really means and how the quality of something is to be measured and expressed. This is less true, of course, when applied to things that are concrete. The quality of many manufactured products can be precisely quantified. This results from the fact that they must conform to standards that are strictly enforceable and are precisely quantifiable—e.g., steel either meets a standard relating to its composition or it does not. In the manufacturing situation, then, "quality control" is not a nebulous idea—it relates to the extent to which products meet the required standards.

In less concrete settings, such as those relating to various types of services, quality is less easily defined. For example, we may refer to the "quality of law enforcement" or the "quality of library service," but these are notions that are more subjective than objective.

Despite it being an imprecise idea in many contexts, it is obvious that the last decade or so has brought a great increase in concern for "quality" in virtually all areas of human endeavor. The growth of the literature on the subject is a tangible manifestation of this.

Nevertheless, it is somewhat misleading to speak of quality as though it were a single idea. Instead, one may recognize various levels or perspectives, as illustrated in Table 1. At the one extreme, there is the abstract or transcendental idea of quality, one that is static, absolute, and existing only in philosophical and metaphysical speculation. At the other extreme is the "user" perspective, which is personal and even, perhaps, idiosyncratic.

It is also dynamic and "relative"—in the sense that it often involves a comparison and the choice of one among several alternatives. Frequently the choice will be made on the basis of cost, which could be a cost in monetary form or in terms of time and convenience.

Table 1.
VARIOUS POSSIBLE LEVELS OR PERSPECTIVES RELATING TO QUALITY

Perspective	*Basis of judgment*	*Characteristics*
ABSTRACT	Philosophy Speculation	Absolute Static
ORGANIZATIONAL		
PROCESS	Standards, regulations, norms	Some processes may be strictly regulated; others not
PRODUCT	Standards	For manufactured products, may be objective and enforceable
SERVICE	Standards or norms	More subjective than objective; rarely enforceable
USER/CUSTOMER	Cost Value Personal value system	Dynamic Relative

Between these extremes, we have other levels or perspectives, identified in the table as being "organizational." Quality related to products varies greatly with type of product. For the many products that must be manufactured to conform to standards, quality can be considered close to absolute, at least relative to the standards, but not completely so since most manufacturing standards accept a range of values, albeit a very narrow one in many cases. Intellectual products, such as various forms of publication, are less susceptible to true standardization. At least, this is true of their content. The container (paper, binding, and so on) can be standardized.

The process perspective is heterogeneous. Some processes can be standardized. In fact, in some cases, processes may be subjected to absolute regulation—e.g., concerning cleanliness, safety, and other health-related issues. Again, intellectual processes are not as susceptible to regulation or standardization.

The service perspective falls midway between the product perspective and the user perspective. Services can rarely be judged in absolute terms. Although some aspects of service can be quantified—e.g., number of seats per reader, number of students per instructor—the standards are

rarely completely enforceable so they tend to be normative values rather than true standards, and some services (e.g., associated with organized religion or with certain social agencies) seem not susceptible to evaluation against any type of standard.

Nevertheless, approaches to the enforcement of quality within service agencies have become increasingly sophisticated in the last several years, culminating in adoption of the principles of total quality management (TQM), which include emphasis on customer satisfaction and on continuous improvement.

Quality in Information Service Settings

Since information tends to be intangible, it is quite difficult to obtain agreement on appropriate measures of quality for most elements of information service. All of the various perspectives represented in Table 1, except for the purely philosophical, can apply in the information service environment. Quality can be considered in tangible terms for many aspects of information products but can be quite elusive elsewhere, especially in both the service perspective and the user perspective. Take, for example, the case of an electronic database. Quantifiable measures of quality can be applied when the database is considered as a *product*—i.e., its coverage of the literature within its scope, the average number of access points per item, up-to-dateness, and so on. Retrospective search and current awareness services derived from use of the database present more difficult problems. While certain measures of service quality can be objective and quantified (e.g., average time elapsing from demand to delivery of response), the more important measures, such as those of recall and precision, are both subjective and difficult to apply. When the user perspective is considered here, of course, the situation becomes even more subjective. For example, a database search can retrieve many items that match a user's stated request or stored interest profile but may still be judged of little value by the user, because the actual information needed did not appear in the search results, because the items retrieved were already known to the user, because he considered them as insignificant contributions to the subject, or for some other reason that might be quite idiosyncratic. Moreover, if the user has to pay for the service, he may apply a purely cost-effectiveness measure to judge the quality of the search results—i.e., the cost per useful item retrieved.

The process perspective on quality is not as nebulous as the user perspective, but it is still an area in which it is difficult to apply true standards. This is because many of the processes are intellectual. While certain applications can be standardized (e.g., form of name in catalog entries), others, such as subject indexing, are not susceptible to standardization except in very trivial aspects. Quality concerns applied to another intellectual process, abstracting, is the focus of our present discussion.

Quality Considerations Applied to Abstracting

From a psycholinguistic perspective, abstracting is more ambitious and complex than indexing: not only must the text of documents be analyzed in some detail but text (the abstract) must also be produced. This text must be coherent syntactically and semantically and, at the same time, be a reasonable summary of the original document. Abstracting is the most difficult of all operations normally applied in a document processing environment because, today at least, an abstract must act as both content description and retrieval tool. Fidel (1986) has shown that these two uses may not be completely compatible.

A possible model of the abstracting process is presented in Figure 1. In actual fact, four levels of processing are represented. The goals are defined by the service or journal producing the abstracts and may be embodied or reflected in guidelines for the abstractors. The individual abstractor observes the goals by following these guidelines. The two processes, "content interpretation/selection" and "content transformation," are directly equivalent to the conceptual analysis and translation stages of subject indexing (Lancaster, 1998). The former is concerned with understanding what is discussed in the original text and deciding which elements should be included in the abstract, while the latter is concerned with the composition of the abstract—i.e., how the selected elements are to be presented in the text of the abstract.

The process headed "checking" is the process directly related to quality. It has several possible dimensions: the individual abstractor may impose his/her own review of quality before submitting the abstract for further processing, the abstractor's work may later be checked by an editor or senior abstractor before publication, and readers may apply their own quality checks relating to the intelligibility of the abstract and its value in predicting the relevance of the original item to their own interests.

Figure 1 suggests that the quality of the abstract is largely determined by the quality of the knowledge base of the abstractor. The knowledge base incorporates both linguistic knowledge (ability to interpret the language of texts in the subject area dealt with) and nonlinguistic knowledge: understanding of the subject matter, of the needs and interests of the audience served, and of the guidelines under which the abstractor is to operate.

Despite the fact that their application in retrieval (as substitutes for or complements to sets of index terms) makes them more important now than ever before, especially in the Internet environment (Wheatley & Armstrong, 1997), there exist no generally accepted measures of the quality of abstracts. Of course, many writers have identified their desirable attributes. Borko and Bernier (1975), for example, regard abstracting as a form of writing that has a unique style (it is not a "natural" form); abstracts must be brief, accurate, and clearly written. Unlike Cremmins

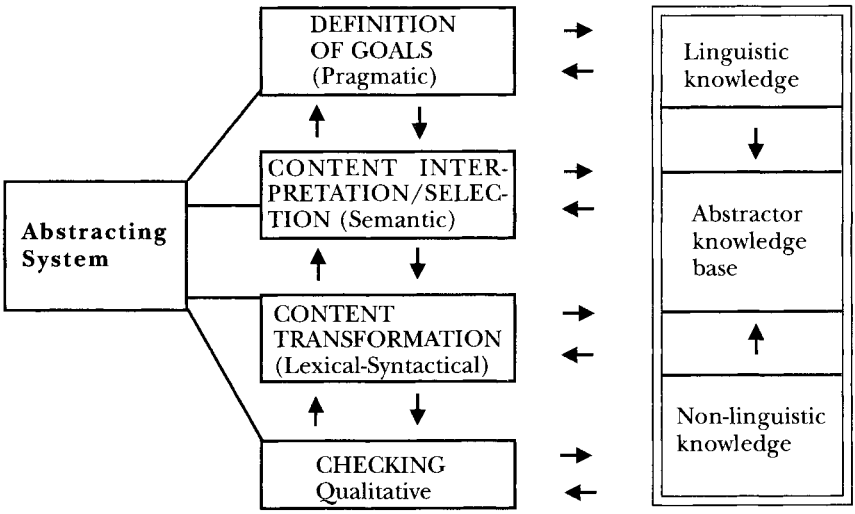


Figure 1. Integrated Model of the Abstracting Process.

(1996), they do not claim that they must have "elegance." Lancaster (1998) suggests two broad criteria for judging quality: are the major points of the article covered and are they represented accurately, succinctly, and unambiguously? The latest English-language standard (National Information Standards Organization, 1997), while it gives guidance on style, makes no attempt to provide criteria that can be used to assess quality. Other writers (e.g., Brown & Day, 1983) have focused on the art of text summarization or on the skills needed by a good abstractor (e.g., see Endres-Niggemeyer, Maier, & Sigel, 1995).

Interest in the evaluation of abstracts can be traced back to at least the late 1950s. For example, Edmundson et al. (1959) proposed several criteria: comparison with an "ideal" abstract, the retrievability of a document by the abstract, and the extent to which the abstract could be used to answer test questions as well as the use of intuitive subjective judgment. Payne, Munger, and Altman (1962) also suggested a test of the value of abstracts in answering questions, as well as a measure of the amount of text reduction achieved in an abstract, and the use of a consistency test in which the similarity of different abstracts, prepared from the same document, is compared. Vinsonhaler (1966) recommended use of a seven-point scale to determine the similarity between an abstract and the document it relates to; also proposed was a more conventional approach, one of predictive validity—the extent to which abstracts are able to correctly predict the relevance of documents.

Mathis (1972) offered a numerical value, known as the "data coefficient" (DC), for the evaluation, expressed by a formula that incorporates

a data retention factor and a length retention factor. The value of the DC is increased by reducing the number of words in the abstract, by increasing the number of concepts ("data elements") represented, or both.

Several of these approaches have been applied over the years. The most favored is a test of the ability of an abstract to predict the relevance of a document to a particular information need. Investigators who have applied this to abstracts, or to extracts derived by computer, include Rath, Resnick, and Savage (1961); Resnick (1961); Kent et al. (1967); Dym (1967); Shirey and Kurfeerst (1967); Saracevic (1969); Marcus, Benenfeld and Kugel (1971); Thompson (1973); and Keen (1976).

Hartley, Sydes, and Blurton (1996) provide an example of a study in which abstracts are judged on their ability to answer various questions; in this case, they were comparing "structured" abstracts with unstructured ones. Salton et al. (1997) used a variation of the similarity approach: the extent to which an automatically-derived extract resembles one derived by humans.

Other approaches have assessed the "readability" of abstracts using standard readability formulas, comprehension measures, or both. Examples can be found in the work of Dronberger and Kowitz (1975), King (1976), Tenopir and Jacsó (1993), and Hartley (1994). More recently, Wheatley and Armstrong (1997) studied readability of a variety of abstracts drawn from Internet sources.

A more "linguistic" approach was used by Salager-Meyer (1991), who analyzed a sample of medical abstracts from this perspective, finding almost half to be "poorly structured" (i.e., having discoursal deficiency). Since "discoursal deficiency" can include such things as conceptual scatter (e.g., results reported in different places in the abstract), as well as omission of an important element (e.g., purpose of research) from the abstract, the author implies that abstracts flawed in this way will be less effective in conveying information. Elsewhere, Pinto (1992, 1994, 1995) has dealt in detail with the process of text summarization from the viewpoint of linguistic structure.

It is clear that the various quality criteria proposed or used in the past look at abstracts/abstracting from different perspectives. In fact, virtually all perspectives represented in Table 1 can apply to abstracts or abstracting, as shown in Table 2.

The process perspective deals primarily with attributes of cognitive representation. Here analogies can be drawn between the process of abstracting and the process of indexing (Lancaster, 1998). The exhaustivity of the abstract relates to its breadth of coverage. In essence, it is a measure of the extent to which all of the themes of the original text are represented in the abstract. Clearly, an abstract is unlikely to include all the content of the original text (unless it is completely trivial) so the exhaustivity of the abstract can be considered as the extent to which all of

Table 2.
ATTRIBUTES OF QUALITY ASSOCIATED WITH DIFFERENT PERSPECTIVES ON ABSTRACTS AND ABSTRACTING

Process perspective	*Service perspective*
Exhaustivity	Customer satisfaction
Accuracy	Cost-effectiveness
Readability	
Cohesion/coherence	*User perspective*
Cost	Cost
	Value
Product perspective	
Consistency	*Process/product perspective*
Brevity	Density
Cost	Cost

the themes (ideas, conclusions, or whatever) judged important are covered in the abstract. This implies that some group of people, presumably specialists in the subject area dealt with, can agree on what is important in the original and what is not.

In an ideal situation, of course, an abstract should be tailored to the needs of a particular audience. This is most obvious in the case of one written for an in-house bulletin prepared, for example, to serve a particular company or research organization. In this case, an exhaustive abstract would be one that covers all the themes of the original that are of potential interest to the limited community. In an extreme case, this might be a single theme—e.g., results of applying a particular drug extracted from a medical article discussing multiple approaches to the treatment of some disease. Clearly, the writer of such an abstract must have a good knowledge of the needs and interests of the target community as well as familiarity with the subject matter dealt with. The more heterogeneous the interests of the audience served, the less likely one is to reach agreement on which themes to include in the abstract and which not: difficult in the case of general mission-oriented abstracts (e.g., serving the needs of an entire industry), more difficult still in the case of abstracts intended to serve the needs of an entire discipline.

Accuracy refers to the extent to which the abstract correctly represents the original text. A theme covered in the abstract could be an inaccurate representation of the original because of an intellectual error (the abstractor misinterprets the text) or an error of carelessness (the abstractor records incorrectly—e.g., gives a wrong numerical value). The former should be relatively rare but could occur if the abstractor is not fully familiar with the subject matter or if the original text is somewhat obscure. A special case would be the situation of an abstractor dealing with a language in which he is not completely fluent. Accuracy errors of the second type would be attributable to personal characteristics of the abstractor (ability to concentrate, ability to transcribe correctly), including qualities

that could vary considerably from one day to the next, and to working conditions. Most significant of the latter would be pressures associated with required productivity, where an abstractor may be required to produce a specified number of abstracts in a particular time period. Of course, once the abstract has been printed and distributed, it would be impossible to determine whether an error of this type was attributable to the abstractor or was introduced at some later stage of the production process.

The readability of an abstract is determined by the ability of the abstractor to express himself clearly, concisely, and unambiguously, by the rules or guidelines under which he operates, and by the format of the abstract (e.g., some claim that abstracts structured into paragraphs with topical headings are easier to comprehend). To the extent that general tests of the readability of text (e.g., the Flesch Reading Ease formula) or of comprehension (e.g., cloze criteria) are applicable to abstracts, readability can be an objective measure and one that can be quantified.

Cohesion/coherence is related to readability but is not identical with it. These properties relate to connectivity between different parts of a text. Extracts prepared by computer (selecting sentences on the basis of statistical, positional, or linguistic criteria) will frequently be lacking in these properties, even though the total extract may be a satisfactory representation of the principal themes of the original text. Salager-Meyer (1991) is perhaps the only author to apply such linguistic criteria to humanly prepared abstracts. A major measure used was that of conceptual scatter—the extent to which related elements (e.g., results) are separated in an abstract. Since structured abstracts (see Haynes, 1993; Hartley, 1994; Hartley, Sydes, & Blurton, 1996) are formatted into paragraphs with preestablished subheads (e.g., methods, results), they are less likely to exhibit such conceptual scatter. Factors affecting cohesion/coherence are the same as those affecting readability.

The product perspective (see Table 2) relates to the technical adequacy of the abstract. The idea of *consistency* in abstracting is similar to consistency in subject indexing. It refers to the degree to which two individuals produce abstracts that are similar to each other (interabstractor consistency) or the degree to which the same individual agrees with himself when abstracting a document on different occasions (intra-abstractor consistency). In the indexing situation, a distinction can be made between consistency in conceptual analysis and consistency in the translation of the conceptual analysis into a particular vocabulary (e.g., terms drawn from a thesaurus). Consistency in abstracting, however, applies only at the conceptual level since it is unrealistic to expect different individuals to use exactly the same words or grammatical constructions. Presumably, consistency will be greatest when abstractors work to precise rules as to what to include and what not. For obvious reasons, structured abstracts should be more consistent than others.

In abstracting, just as in indexing, consistency is not the same as quality (Cooper, 1969). Nevertheless, if two abstractors (or indexers) consistently produce similar results, while a third agrees little with the other two, one is generally inclined to believe that the consistent abstracting (indexing) will be "better." Salton, Singhal, Mitra, and Buckley (1997) justify their automatic procedures for selecting and linking pieces of text on the grounds that the summary thus produced is as likely to agree with a humanly-produced summary as one humanly-produced summary is to agree with another. In translating from one language to another also, consistency (similarities) has been suggested as an indicator of quality (Brew & Thompson, 1994).

Brevity is an obviously desirable attribute of a good abstract, and it is susceptible to exact measurement. Moreover, length is one of the few attributes that the published standards can and do address precisely, at least in terms of a recommended range in number of words. Nevertheless, brevity should always be secondary to other considerations such as exhaustivity and accuracy. Moreover, absolute standards make little sense since several factors would influence the brevity: length, complexity or diversity of the original, type of abstract (indicative, informative, critical), and accessibility of the original (one could argue that materials less physically or intellectually accessible—e.g., published in obscure sources or unfamiliar languages—should be abstracted more fully).

Cost can be related to abstracts at different levels: the intellectual cost of creating an abstract, the cost per abstract of producing a printed publication, the cost per abstract in distribution (e.g., as part of a current awareness service), and so on. Factors affecting cost differ from level to level. For example, abstract length has a major effect on the cost of producing a printed publication but much less effect on the inclusion of an abstract in an electronic database. Cost of writing the abstract in the first place depends most obviously on who the writer is, how much he/she is paid, and who is paying. The cost of abstracting can be looked at from several different perspectives. For example, use of author-generated abstracts is economical for database producers. From the much broader (society) perspective, however, they are very expensive since the time of such authors as research scientists can be considered to be so valuable that it is perhaps better spent on other things.

Carried to its logical conclusion, of course, one could argue that the greatest cost associated with abstracting is the cost of the time spent by people in reading the abstracts (thus the importance of such factors as brevity and readability) and in taking actions based upon them (thus the importance of such factors as accuracy and exhaustivity). Cost, then, is a multifaceted attribute when related to abstracts and abstracting. For this reason, it appears within all the perspectives illustrated in Table 2.

Density is a measure that relates the attribute of exhaustivity to that of brevity. It thus, in a sense, combines the process and product perspectives. Given that the abstract includes everything that should be included—all the topics of potential interest to the intended audience—the briefer the abstract the better providing, of course, that other requirements, such as readability, are not significantly degraded. Density, then, refers to the amount of information content provided by an abstract of a certain length. The density of an abstract can be considered related to its entropy—the extent to which uncertainty about the original document is reduced for the reader of the abstract. Standard tests of the relevance predictability of abstracts address this issue.

The data coefficient proposed and tested by Mathis (1972) was a precise measure of density, defined by the equation DC = C/L—i.e., the data coefficient (DC) is the "data retention factor," C, divided by the "length retention factor," L. The C value is the measure of exhaustivity as defined earlier in this discussion, while the L value is the number of words in the abstract divided by the number in the original. Clearly, the DC of an abstract improves as either exhaustivity or brevity increase.

While the process and product perspectives consider abstracts as entities in their own right, the service perspective is obviously concerned with their application. Providers of abstracts, whether publishers and editors of scholarly journals or producers of secondary databases in printed or electronic form, are presumably concerned with offering a product that the majority of their customers (journal readers, database users) will find acceptable. Customer satisfaction will most obviously be associated with the process and product parameters discussed earlier, perhaps most closely to accuracy, readability, and exhaustivity. Clearly, the providers will also be concerned with production and distribution costs so, ultimately, "quality" becomes a matter of cost-effectiveness—i.e., customer satisfaction at least cost.

As mentioned earlier, the user perspective on quality will tend to be subjective, relative, dynamic and, perhaps, idiosyncratic. Users of abstracts will be likely to judge their quality in practical and pragmatic terms. They are unlikely to demand elegance but they will expect readability. Ultimately, they will judge abstracts and abstracting services in terms of costs and value to themselves. Taking the user's own time into account, the predictive validity of the abstract is of paramount importance. That is, users will be unhappy with a service whose abstracts frequently cause the incurring of costs associated with obtaining complete texts that turn out to be irrelevant. Nor will they be satisfied with one that frequently fails to lead them to sources that they would judge valuable if seen in full form.

Current Methods

The automatic processing of text has increased considerably over the

years as computing power has increased, computing and storage costs have decreased, and more and more text has become available in electronic form, largely as a byproduct of various forms of publishing. The development of the Internet and the World Wide Web, which makes vast quantities of text accessible to huge numbers of users, has made text search the norm rather than the exception. As might be expected from all of this, interest in automatic text processing methods has increased very greatly in the 1990s, in the research community as well as in government and commercial sectors. Current approaches to the processing of text, for information retrieval and related purposes, are well portrayed in the proceedings of a series of conferences. Most important among these have been the Text Retrieval Conferences (TREC) organized by the (U. S.) National Institute of Standards and Technology (Sparck Jones, 1995; Harman, 1997), the Message Understanding Conferences (MUC), the Conferences on Applied Natural Language Processing, and the International Conferences on Document Analysis and Recognition. The TREC and MUC conferences are particularly important for their methodology: all participating research groups must apply their text processing procedures to some common pre-established tasks, allowing performance comparisons across the methods.

Current methods of text processing for information-retrieval-like purposes go beyond text search, automatic indexing and automatic extracting procedures (all of which have existed, to some extent at least, since the late 1950s), now including such activities as text linkage, text augmentation, and text generation. Nevertheless, while current approaches may achieve rather better results, they do not differ much in principle from those first introduced forty to fifty years ago, even though they may be given different names ("text summarization" in place of abstracting/extracting, "text categorization" in place of indexing/classification, and so on) and may be more sophisticated in some respects (e.g., not just extracting text but putting the extracts into a pre-established template). While some current approaches claim to apply techniques drawn from artificial intelligence research, and the term "intelligent text processing" is sometimes used to refer to procedures of this type (see, for example, Jacobs, 1992), it is doubtful that any can be considered to exhibit true intelligence (Lancaster & Smith, 1999).

Knowledge Discovery

The great majority of the criteria of quality proposed and used in the past apply most obviously to abstracts intended to be read by humans. As mentioned earlier, if abstracts are intended primarily as useful document surrogates for search purposes, the quality criteria become somewhat different. Unfortunately, a good abstract for search purposes is unlikely to be good for a human reader. Indeed, an abstract prepared solely for

computer searching, such as the telegraphic abstracts of the semantic code system (Perry & Kent, 1958), may not be readable by humans at all, and abstracts prepared primarily for search purposes, such as the mini-abstracts proposed by Lunin (1967), may be somewhat difficult for humans to comprehend.

For retrieval purposes, and especially in knowledge discovery tasks, exhaustivity and accuracy are extremely important, and the other attributes in Table 2 diminish in significance. In fact, for abstracts intended solely for search purposes, such criteria as readability and coherence/cohesion are not important at all, while other attributes are applicable in opposite ways. Most obviously, brevity is not necessarily desirable since the retrievability of an abstract will be directly related to its length (i.e., number of access points provided). Nevertheless, for reasons mentioned before, there is likely to be an optimum length for effective search and discovery operations. The data retention factor proposed by Mathis (1972) seems a particularly appropriate criterion in knowledge discovery applications since it relates length to completeness of content coverage. Also undesirable for knowledge discovery purposes is internal consistency because redundancy improves retrievability. That is, if a particular idea is expressed in different ways in an abstract (no synonym control), this increases the probability that the text will match an expression selected by a particular searcher or that meaningful relationships between related ideas will be revealed.

Conclusion

Text surrogates for larger bodies of text, whether one refers to them as "abstracts," "summaries," or some other term, have proved extremely useful in a wide variety of information processing applications for very many years. The increasing application of computers to text processing has not reduced their value (although criteria for judging their quality may have changed somewhat), and one has no reason to suppose that their value diminishes as more critical or sophisticated operations, including those of knowledge discovery, are applied to the text.

References

Borko, H., & Bernier, C. L. (1975). *Abstracting concepts and methods.* New York: Academic Press.

Brew, C., & Thompson, H. S. (1994). Automatic evaluation of computer generated text: A progress report on the TextEval project. In *Proceedings of the Human Language Technology Workshop* (March 8-11, 1994) (pp. 108-113). San Francisco, CA: Morgan Kaufmann.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*(1), 1-14.

Cooper, W. S. (1969). Is inter-indexer consistency a hobgoblin? *American Documentation, 20*(3), 268-278.

Cremmins, E. T. (1996). *The art of abstracting,* 2d ed. Arlington, VA: Information Resources Press.

Dronberger, G. B., & Kowitz, G. T. (1975). Abstract readability as a factor in information systems. *Journal of the American Society for Information Science, 26*(2), 108-111.

Dym, E. D. (1967). Relevance predictability: I. Investigation, background and procedures. In A. Kent, O. E. Taulbee, J. Belzer, & G. D. Goldstein(Eds.), *Electronic handling of information: Testing and evaluation* (pp. 175-185). Washington, DC: Thompson Book Co.

Edmundson, H. P.; Oswald, V. A., Jr.; & Wyllys, R. E. (1959). *Automatic indexing and abstracting of the contents of documents.* Los Angeles, CA: Planning Research Corporation.

Endres-Niggemeyer, B.; Maier, E.; & Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing & Management, 31*(5), 631-674.

Fidel, R. (1986). Writing abstracts for free-text searching. *Journal of Documentation, 42*(1), 11-21.

Harman, D. (1997). The TREC conferences. In K. Sparck Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 247-256). San Francisco, CA: Morgan Kaufmann.

Hartley, J. (1994). Three ways to improve the clarity of journal abstracts. *British Journal of Educational Psychology, 64*(1), 331-343.

Hartley, J., & Sydes, M. (1996). Which layout do you prefer? An analysis of readers' preferences for different typographic layouts of structured abstracts. *Journal of Information Science, 22*(1), 27-37.

Hartley, J.; Sydes, M.; & Blurton, A. (1996). Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science, 22*(5), 349-356.

Haynes, R. B. (1993). More informative abstracts: Current status and evaluation. *Journal of Clinical Epidemiology, 46,* 595-597.

Jacobs, P. S. (Ed.). (1992). *Text-based intelligent systems: Current research and practice in information extraction and retrieval.* Hillsdale, NJ: Lawrence Erlbaum.

Keen, E. M. (1976). A retrieval comparison of six published indexes in the field of library and information science. *Unesco Bulletin for Libraries, 30*(1), 26-36.

Kent, A.; Belzer, J.; Kurfeerst, M.; Dym, E. D.; Shirey, D. L.; & Bose, A. (1967). Relevance predictability in information retrieval systems. *Methods of Information in Medicine, 6*(2), 45-51.

King, R. (1976). A comparison of the readability of abstracts with their source documents. *Journal of the American Society for Information Science, 27*(2), 118-121.

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice,* 2d ed. Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Lancaster, F. W., & Smith, L. C. (In press). *Intelligent technologies in library and information service applications: A realistic appraisal.* Medford, NJ: Information Today.

Lunin, L. (1967). The development of a machine-searchable index-abstract and its application to biomedical literature. In B. Flood (Ed.), *Three Drexel information science-research studies* (pp. 47-134). Philadelphia, PA: Drexel Press.

Marcus, R. S.; Benenfeld, A.R.; & Kugel, P. (1971). The user interface for the Intrex retrieval system. In D. E. Walker (Ed.), *Interactive bibliographic search: The user/computer interface* (pp. 159-201). Montvale, NJ: AFIPS Press.

Mathis, B. A. (1972). *Techniques for the evaluation and improvement of computer-produced abstracts.* Columbus: Ohio State University, Computer and Information Science Research Center (OSU-CISRC-TR-72-15. PB 214 675).

National Information Standards Organization. (1997). *Guidelines for abstracts.* Bethesda, MD: NISO.

Payne, D.; Munger, S. J.; & Altman, J. W. (1962). *A textual abstracting technique: A preliminary development and evaluation support.* Pittsburgh, PA: American Institutes for Research (2 vols. AD 285081-285082).

Perry, J. W., & Kent, A. (1958). *Tools for machine literature searching.* New York: Interscience Publishers Inc.

Pinto, M. (1995). Documentary abstracting: Toward a methodological model. *Journal of the American Society for Information Science, 46*(3), 225-234.

Pinto, M. (1994). Interdisciplinary approaches to the concept and practice of written text documentary content analysis (WTDCA). *Journal of Documentation, 50*(2), 111-133.

Pinto, M. (1992). *El resumen documental: Principios y métodos.* Madrid: La Fundación Germán Sánchez Ruipérez.

Rath, G. J.; Resnick, A.; & Savage, T. R. (1961). Comparison of four types of lexical indicators of content. *American Documentation, 12*(2), 126-130.

Resnick, A. (1961). Relative effectiveness of document titles and abstracts for determining relevance of documents. *Science, 134*(3484), 1004-1006.

Salager-Meyer, F. (1991). Medical English abstracts: How well are they structured? *Journal of the American Society for Information Science, 42*(7), 528-531.

Salton, G. (Ed.). (1971). *The SMART retrieval system: Experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall.

Salton, G.; Singhal, A.; Mitra, M.; & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management, 33*(2), 193-207.

Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgements. *Proceedings of the American Society for Information Science, 6,* 293-299.

Shirey, D. L., & Kurfeerst, M. (1967). Relevance predictability: II. Data reduction. In A. Kent; Taulbee, O. E.; Belzer, J.; Goldstein, G. D. (Eds.), *Electronic handling of information: Testing and evaluation* (pp. 187-198). Washington, DC: Thompson Book Co.

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing & Management, 31*(3), 291-314.

Tenopir, C. (1985). Full text database retrieval performance. *Online Review, 9*(2), 149-164.

Tenopir, C., & Jacsó, P. (1993). Quality of abstracts. *Online, 17*(3), 44-55.

Thompson, C. W. N. (1973). The functions of abstracts in the initial screening of technical documents by the user. *Journal of the American Society for Information Science, 24*(4), 270-276.

Vinsonhaler, J. F. (1966). Some behavioral indices of the validity of document abstracts. *Information Storage and Retrieval, 3*(1), 1-11.

Wheatley, A., & Armstrong, C. J. (1997). Metadata, recall, and abstracts: Can abstracts ever be reliable indicators of document value? *Aslib Proceedings, 49*(8), 206-213.

Knowledge Discovery in Spatial Cartographic Information Retrieval

LIXIN YU

ABSTRACT

LIBRARY CATALOGS FOR MAP COLLECTIONS are not well developed in most libraries. The cartographic information source differs from other kinds of information in that it is usually rectangular in shape and defined by the coordinates of the four map corners. This coordinate information proves difficult for many people to use, unless a certain user interface is designed and knowledge discovery algorithms are implemented. A system with such an interface and algorithms can perform powerful queries that an ordinary text-based information retrieval system cannot. This article describes a prototype system—GeoMatch—which allows users to interactively define geographic areas of interest on a background map. It also allows users to define, qualitatively or quantitatively, the relationship between the user-defined area and the map coverage. The knowledge discovery in database (KDD) factor is analyzed in the retrieval process. Three librarians were interviewed to study the feasibility of the new system. The MARC record format is also discussed to argue that conversion of cartographic material records from an existing library online catalog system to GeoMatch can be done automatically.

INTRODUCTION

Knowledge discovery in databases (KDD) has become a hot topic in recent years. The KDD method has been used in various fields, including spatial database analysis (Xu et al., 1997), automatic classification (Bell, 1998), deviation detection (Schmitz, 1990), and clustering (Cheesman,

Lixin Yu, School of Information Studies, Florida State University, Tallahassee, FL 32306-2100
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 249-263

1996). This article explores the use of KDD in information retrieval by examining the nature and process of geographic information retrieval. It deals with the characteristics of Geographic Information Systems (GIS), Bibliographic Records for Cartographic Information, and a GIS-based cartographic information retrieval system—GeoMatch.

GIS and Functions Related to the GIS-Based Information Retrieval System

The Environmental System Research Institute (ESRI) is the largest GIS software producer in the world. ESRI defines GIS in its menu (Environmental System Research Institute, 1991) as: "An organized collection of computer hardware, software, geographic data, and personnel designed to efficiently capture, store, update, manipulate, analyze, and display all forms of geographically referenced information." Most words in this definition can be found in definitions of many other information systems. What makes GIS special is the term *geographically referenced data.* GIS uses spatial location as the major link to organize and manipulate information.

A typical GIS has two major functional components—a database management system, which stores and manipulates the data, and a spatial engine, which performs special topological operations on geographic features. A common misunderstanding of GIS is to consider it merely a computerized mapmaker. GIS is a powerful analytical tool that is far more sophisticated than a mapmaker. It is true that some GIS products on the market are simplified for naïve GIS users to generate, view, and print maps. These "viewer"/software packages often support only limited data manipulation functions. They are not considered fully functional GIS systems. A GIS can perform network analysis, overlay, buffering, and many other operations that few other information systems can accomplish. As Burrough (1990) summarized, a GIS can answer such questions as:

- Where is 785 S. Allen Street in Albany, New York?
- In what census tract is the above address located?
- How many supermarkets are within three miles from the above address?
- A delivery truck needs to deliver items to 200 customers. What is the shortest route and sequence to make the delivery? If road traffic information is available, what is the fastest route to finish the task?
- Given the population in a county, what is the population density? (GIS can calculate the area of the county precisely).
- A new shopping mall is going to be built in the city. The mall should be built at least five miles away from the existing shopping malls; next to a major street; surrounded by 5,000 residents within four miles; and no more than ten miles from the downtown area. Where is the best place to build the new mall?

There are many other questions that only a GIS can answer. One of the GIS functions that is highly related to the geographic information retrieval system is **overlay**. Some concepts need to be defined to understand the overlay process.

In a GIS, a **polygon** is an enclosed area bounded by lines such as a census tract or a county. Consequently, polygons have areas and parameters that a GIS can calculate. A **layer** or a **theme** is a concept for a *single feature map* in GIS. For example, a county map of Florida showing the average age of a population is a polygon layer. These single-feature layers can be integrated by GIS for analysis.

GIS has the capability of building geometric topology. It can determine which lines are crossing one another to create a node at the cross point. It can detect what lines are connected to create an enclosed polygon. GIS can then generate a polygon object with features like area and parameter. The topology in a GIS can be expressed as the relationship of points, lines, and polygons. GIS can do sophisticated spatial analysis after the topology is established.

The process of merging multiple layers is called overlay, a unique function of GIS. For example, assume that there are two maps printed on transparencies—a map of census tracts and a map of a lake, all in the same county. If both maps are in exactly the same scale and the four corners of the two maps represent exactly the same locations, the two transparencies can be put together to make a new map—with both county boundaries and the lake shore. The new map is the so-called overlay. GIS is very powerful in performing this operation. It can overlay maps with different kinds of features (point, line, polygon) and develop new topologies for further analysis. Burrough (1990) lists forty-four kinds of overlay analysis capabilities that GIS may have. Figure 1 demonstrates the overlay process. The first map layer shows school district boundaries(District C and District D). The second map layer represents county boundaries (County A and County B). During the overlay process, GIS combines the features from both map layers into a third layer that contains four polygons. In the third map layer, each polygon will have attributes from both the county map layer and the school district map layer. For example, area 1 will have its area, parameter, county name A, school district name C, and other data previously stored in the two map layers. Obviously, it would be difficult to integrate the school district data and county data like this using only database techniques because the data collected represent different areas.

Knowledge Discovery in Databases and Information Retrieval

Due to the less expensive data storage and increasing computing power, the volume of data collected by various organizations has expanded

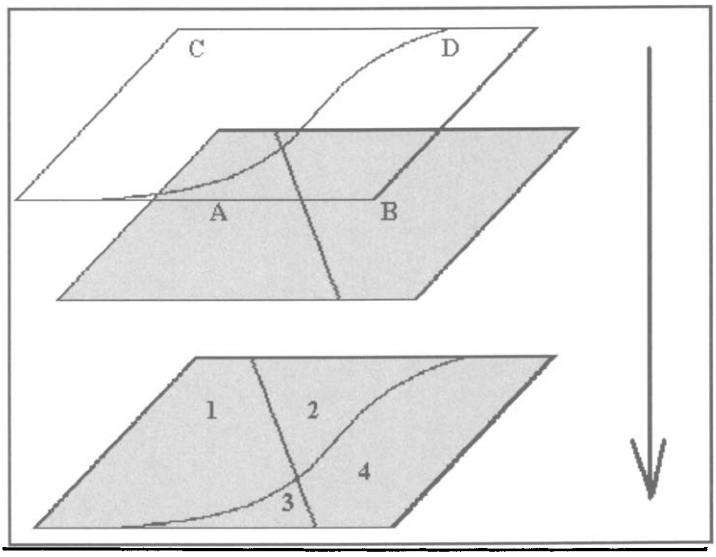


Figure 1. Overlay Process.

rapidly. This vast abundance of data, often stored in separate data sets, makes it more difficult to find relevant information. On the other hand, the power of computers also makes it possible to integrate the data sets, compile the facts, and develop the information into "a collection of related inferences" (Trybula, 1997). This is why KDD has received such attention from both the academic and commercial worlds. According to Tuzhilin (1997), the number of papers submitted to the Knowledge Discovery Workshop increased from 40 in 1993 to 215 in 1996.

Fayyad, Piatetsky-Shapiro, and Smyth (1996) define KDD as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns of data" (p. 2). As Trybula (1997) summarized, the methods of evaluating data include algorithms, association, change and deviation determination, visualization, and sixteen other analytical techniques. No matter which method is employed, the key point of KDD is to uncover new, useful, and understandable knowledge.

Information retrieval can be simply expressed as a matching process—matching a user's information need with the information source (School of Information Studies, 1998). In this process, a user must express his/her information need accurately so that the system can retrieve the information. On the other hand, information sources need to be organized in such a way that the most important attributes, such as title, author, subject terms, keywords, publication year, and so on, are readily available.

Text information retrieval systems have become more powerful in the last three decades. The retrieval efficiency and effectiveness has been greatly improved through Boolean operators, truncations, proximity, probability search, and many other search mechanisms. However, some attributes in bibliographic records can create difficulty for exact match in a search. Some attributes are even difficult for users to understand. For example, geographic coordinates are attributes in MARC records for cartographic data. Few users would want or be able to enter exact numbers to match those coordinates. Even fewer would know what the numbers mean. Despite these difficulties, however, could the coordinates be useful in information retrieval? Can they be processed to provide understandable and useful knowledge in selecting relevant information?

This article will demonstrate a prototype of a GIS-based cartographic information retrieval system and illustrate how such a system could indeed generate new and useful knowledge during the retrieval process.

Cartographic Information Retrieval

Cartographic Information Retrieval in Libraries

An access point is defined as "a name, term, code, etc., under which a bibliographic record may be searched and identified" (*Glossary*, 1995). An ordinary information retrieval system usually has common access points such as author, title, keywords, subject headings, classification number, and information from other special fields.

In addition to its spatial coverage, a cartographic information source, such as a single sheet map, shares most of the attributes other information sources have, including title and subject terms. A cartographic information source is different from other formats in that, as an information container, it is usually in the shape of a rectangle and contains the coordinates of the four map corners. Nevertheless, most current retrieval systems do not use geographic coordinates as access points because this does not make sense in a text information retrieval system. Many libraries are still in the process of retrospective conversion from card catalogs to text-based online catalogs for their map collections. To study the feasibility of libraries adopting a GIS-based cartographic information retrieval system, long interviews with three librarians were conducted in two libraries in Tallahassee, Florida.

During each interview, a prototype of a GIS-based cartographic information retrieval system (GeoMatch) was demonstrated. The librarians were asked to answer questions concerning the library's map collection, user needs, retrieval tools, and searching procedures. The librarians were also asked to evaluate the usability of the prototype software and assess the usefulness of the system.

Florida State Library

Most of the map collection in the Florida State Library consists of historical maps. Although the library is currently outsourcing the map cataloging to an organization associated with OCLC, the card catalog is still the major retrieval tool for the map collection. The library has added only 800 maps to its online catalog. The online catalog features keyword searching, which provides more retrieval power than the card catalog. The card catalog allows searching only from author, title, and subject terms. During the interviews, the librarians indicated that they had seen more patrons using the catalog since the online version was implemented.

The library has no plan yet to digitize (scan) the maps. Patrons usually cannot find needed maps using the card catalog. Some patrons can locate their maps using the online catalog with keyword searching. Generally speaking, patrons primarily rely on the map librarians to find and access maps.

Although the online catalog system cannot provide sufficient assistance for accessing cartographic information, every day many map users do search historic maps, railroad maps, and place names. Great reliance must be placed on the knowledge and expertise of the map librarians.

Florida State University Library

The Florida State University (FSU) library has a collection of 165,000 single sheet maps, including U. S. Geological Survey maps, road maps, city maps, thematic maps, and historical maps. Records for most of the single sheet maps are maintained in the card catalog. The librarians have started the retrospective conversion of map card catalog records to online catalog records using OCLC. According to the map librarian, most of the records can be found in the OCLC database. During the conversion process, the librarian must make minor changes before adding the OCLC records to the library's online catalog.

The librarians serve many map users everyday including faculty, students, and users referred by other libraries. The map librarians are very familiar with the map collection and usually can find the maps needed. The situation at the FSU library is similar to the one at the Florida State Library—i.e., the map librarians are the most valuable source of information, given the fact that the catalog system for the cartographic data is not very helpful.

In summary, map librarians in both libraries are the most important sources of information for users seeking cartographic data.

Both libraries are in the process of converting cartographic records in the card catalog to the online catalog. The online catalog with searching capability has led to increased map use.

Although most users can access the map information they need with the help of librarians, this situation needs to be improved, for several reasons. First, the map librarians are not certain whether or not they

actually find the maps that best match users' needs. Second, none of the librarians think they can provide a complete list of maps that users might be interested in, especially in a library with more than 100,000 maps. Finally, searching for the right information in such a system relies extensively on human expertise. As one librarian said: "It is at the librarian's mercy whether the user can get a satisfactory answer." If current map librarians leave their positions, it would take new map librarians years to familiarize themselves with the library collection. There exists a great demand for a powerful searching tool for the library map collection.

Studies of Geo-Based Retrieval Tools

A literature review indicates that more advanced cartographic information retrieval systems, designed for searching electronic maps, have been created and are still in the process of refinement. The Alexandria Project is probably the most well-known electronic library system dealing with topological relationships.

Smith (1996) described the goal of the Alexandria Project Digital Library (ADL) as "to build a distributed digital library (DL) for geographically-referenced materials. A central function of ADL is to provide users with access to a large range of digital materials, ranging from maps and images to text to multimedia, in terms of geographical reference" (http://www.dlib.org/dlib.org/dlib/march96/briefings/smith/03smith.html).

The Alexandria Atlas Subteam investigates "the design and functionality of an atlas that would support graphical/geographical access to library materials" (http://wwww.alexandria.ucsb.edu/public-documents/annual-report97/node28.html#SECTION00051300000000000000). As the Alexandria Web site indicates, "spatial searching has not been an available service to library clients and it is not at all clear how ADL clients will react to having actual spatial data available over the Web" (http://www.alexandria.ucsb.edu/public-documents/annual-report97/node28.html#SECTION00051300000000000000). The team is studying such issues as scale, data registration, search result presentation, and fuzzy footprints.

The Alexandria system supports geographical browsing and retrieval using a graphical map interface. An example of the interface can be found at <http://www.dlib.org/dlib/march96/briefings/smith/03smith.html>. Users can zoom in and zoom out on the current view of the map. They can select the map features they wish to see on the background map such as borders and rivers. Users can also select an area of interest and a mode of either OVERLAPS or CONTAINS. An overview of the system is available at <http://www.alexandria.ucsb.edu/adljigi/tutorials/walkthrough1/walkthrou>.

The prototype of GeoMatch has some new functions in addition to those available in the Alexandria system. The initiative of testing GeoMatch is to answer the following two questions: (1) can a GIS/Graphic-based retrieval tool like the Alexandria project be used for nonelectronic cartographic collections in libraries? and (2) what new functions can be developed to improve the GIS-based retrieval tool?

GEO-MATCH—A RETRIEVAL TOOL THAT SEARCHES

Figure 2 illustrates a query screen of the Geo-Match system. In addition to specifying ordinary information needs such as year, title, publisher, keyword, and so on, this system allows a user to interactively identify the interested area using a mouse. It also asks the user to specify the topological relationship between the map coverage and the user-selected area. The system accepts containment and overlapping relationships as summarized by Cobb and Petry (1998). There are two possible containment relationships—the user-selected area falls entirely within a map coverage or the coverage of a map falls within the user-selected area. Users can make a selection.

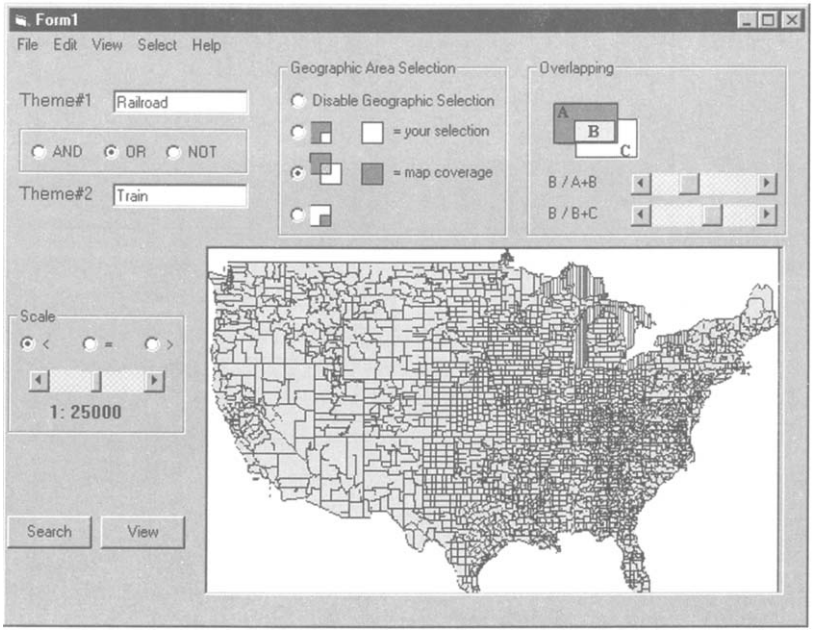


Figure 2. Screen Example of Geo-Match System.

If a user decides to select the overlapping relationship, more choices become available to specify quantitatively the degree of overlap. This degree includes the percentage of the overlapping area in maps and the percentage of the overlapping area in the user-selected area. If a user selects 85 percent as the overlapping criterion in the user-selected area, the user will find maps that cover most of the area of interest (Figure 3). If a user selects 85 percent as the overlapping criterion in the map coverage, the user will find maps that concentrate on the selected area (Figure 4). Users can specify how searching results should be ranked based on the degree of overlap.

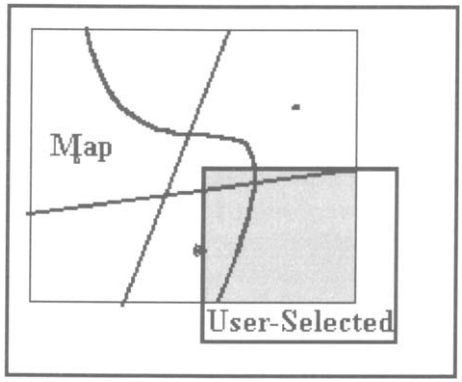


Figure 3. The shaded overlapping area should cover at least 85 percent of the user selected area.

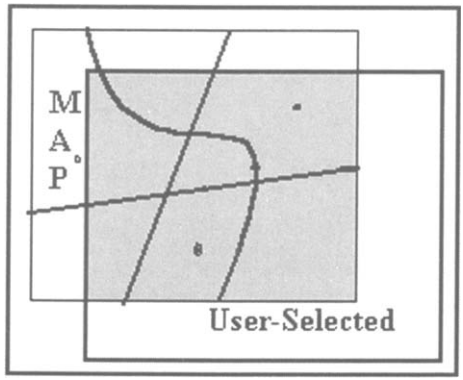


Figure 4 . The shaded overlapping area should cover at least 85 percent of the map

The key features of the prototype are its capability for the user to interactively identify the area of interest—i.e., to quantitatively specify the relationship between the user-defined area and the map coverage, and to rank the search results based on the degree of overlapping.

Use of Graphics to Express Information Need

Cartographic information is geographically referenced—it represents locations and areas on the earth. Conventional information representation using text and symbols is not very useful in describing the information included in a map; there are too many geographic features included in an area. For example, a railroad map in Florida can be indexed using the keywords *railroad* and *Florida.* However, the map also includes all the railroads in each county in Florida. It indicates railroad construction in the Jacksonville area and demonstrates the railroad near Lake xxx. It is practically impossible to index all the place names included in an area. When a user draws a box to specify an area of interest, the information requested would require many words to describe it. A graphic interface can hide the coordinate numbers and present them in scalable graphics, which makes it much easier for users to discover the cartographic information resources of interest.

In addition to the information representation issue discussed earlier, a graphic interface also avoids trouble for users when changes in place names and county boundaries occur or when they simply do not know the exact name to begin the search.

Level 1 in KD—Specifying Topological Relationships Qualitatively between the User-Defined Area and the Map Coverage

As discussed earlier, the Alexandria Project can specify topological relationships qualitatively between the user-defined area and the map coverage in its electronic cartographic information retrieval system. This matching process goes beyond the exact matching in a conventional information retrieval system. The computer system will calculate the topological relationship between the user-defined area and the coverage of the maps to determine whether they overlap or one completely contains another.

Cobb and Petry (1998) presented a model for defining and representing binary topological and directional relationships between two-dimensional objects. Such relationships can be used for fuzzy querying. Cobb and Petry (1998) summarize that there are four kinds of major relationships—disjoint, tangent (next to each other), overlapping, and containment. The assumption for GeoMatch is that users would find overlapping and containment most useful when querying the system.

The operations involved in the above include conversion from screen coordinates to the real world coordinates and comparison of the coordi-

nates of the corners of the user-defined area and map boundaries. The new knowledge—whether two areas overlap—is generated in this process. The knowledge acquired can be utilized to lead users to the relevant information source. GeoMatch provides users with an additional choice beyond the Alexandria system with which to define the containment relationship.

Level 2 in KD—Specifying a Topological Relationship Quantitatively between the User-Defined Area (Rectangle) and the Map Coverage

Specifying a topological relationship quantitatively between the user-defined area and the map coverage is a unique feature of the GeoMatch system. In this process, not only is the topological relationship of the two areas determined, more mathematical calculation is performed to estimate how much the two areas overlap. By combining the information input by users and the data stored in the database, the computer algorithm discovers new knowledge not explicitly represented in the database. Since the user-defined area is rectangular, the calculation involved is not overwhelming and can be realized using a conventional programming language such as C++ or Visual Basic.

This feature allows the system to achieve a higher recall and precision than those systems without this function. Gluck (1995) made an analysis of the relevance and competence in evaluating the performance of information systems. He indicated that "relevance judgments by users most often assess the qualities of retrieved materials item by item at a particular point in time and within a particular user context" (p. 447). Using the qualitative topological matching technique described in Level 1 above, there could be a large gap between the relevance of the system's view and the relevance of the user's view. For example, users may find that some retrieved maps cover only a small part of the area of interest and in fact are useless, but these maps are relevant from the system's view since they overlap the user-defined area. Users may also find that some retrieved maps cover such a large area that the area of actual interest encompasses only a small portion of the whole map. These maps are relevant too from the system's view but, again, practically useless for users. The reason for such a gap between the user's view and system's view is that not enough "knowledge" is discovered and provided for users to describe their information need in more detail. The techniques employed in the quantitative topological matching can greatly reduce the gap of relevance between the two perspectives. In addition, Geomatch can calculate the spatial relevance of the maps to the area of interest and rank the results using the quantitative overlapping factor, while many systems fail to "provide useful ordering of retrieved records" (Larson, McDonough, O'Leary, Kuntz, & Moon, 1996, p. 556). This function is particularly helpful for users when hundreds of maps are included in the result set.

Level 3 in KD—Specifying Topological Relationship Quantitatively between User-Defined Area (Free Style) and Map Coverage

Specifying a topological relationship quantitatively between a user-defined area and map coverage differs from level 2 in that users are allowed to use the mouse to define an irregular area of interest rather than a straight rectangle. This feature can help users express their information need more precisely. For example, a user interested in the lake shore area of a lake can draw an irregular circle around the lake and perform a search.

This process involves complicated topological calculations that are difficult to accomplish using conventional programming languages. The GIS overlay function introduced at the beginning of this discussion needs to be used to generate new polygons and calculate the areas involved. Although the GeoMatch prototype currently does not have this feature, this function could be implemented using a third party GIS software such as the Spatial Engine from ESRI.

MARC Record for Cartographic Information Resources

Whether an information system can be adopted depends not only on its creativity and usefulness but also on the degree of difficulty in converting the current system to the new system. MARC record format is studied to examine what new information needs to be collected to use GeoMatch.

US MARC (Machine Readable Cataloging), developed by the Library of Congress, follows the national standard (ANSI/NISO Z39.50) and international standard. It is the basic format of bibliographic description in the United States. Most online catalogs have a MARC interface for data import and export. OCLC, the bibliographic utility, also provides records in MARC format for members to share.

The current MARC format provides sufficient geographic information to support a more powerful searching tool such as GeoMatch. The most important field is Field 034—Coded Mathematical Data Area Field (Mangan, 1984). If a single set of scales is used, the first indicator is set to "1." The subfield codes include $b (ratio linear horizontal scale); $c (ratio linear vertical scale); $d (coordinates—westernmost longitude); $e (coordinates—easternmost longitude); $f (coordinates—northernmost latitude); and $g (coordinates—southernmost latitude). The following is an example of the MARC record 034 field:

> 034 1 a $b 7603200 $d W1640000 $e W0440000 $f N0900000 $g N040000

The field above illustrates that the map covers an area from West

164°00'00" to West 044°00'00" in longitude and from North 090°00'00" to North 040°00'00" in latitude. This demonstrates that MARC records are capable of defining the scope of a map, and the data are usable in systems like GeoMatch. No additional value-adding operations are necessary unless the bibliographic record of a map is not available from the OCLC database or no matching MARC record is available for the map. If a library already has its map collection in its online catalog, all the records can be imported into GeoMatch automatically.

Feedback from Librarians

Florida State Library

When librarians at the Florida State Library reviewed the prototype for GeoMatch, they realized that it could give answers to difficult questions. For example, towns may disappear over time, county boundaries may change, and users might not remember an exact place name. In such cases, GeoMatch could be very helpful.

Florida State University Library

The librarian showed interest in the GeoMatch system. She thought the system could be useful but should be integrated with the university library catalog system. When the librarian was asked whether the GeoMatch system could solve some difficult to answer questions, she provided the following example:

> Case Study—a man born in 1907 wanted to find a map of his place of birth. He knew the name of the town and knew that it was located west of Jacksonville. He could not find his town on a current map since it has disappeared. He had no idea how to find a map showing the exact position of that town using the library catalog.

In summary, librarians in both libraries confirmed the need for a retrieval tool with a graphic user interface facilitating location-based searching. Such a tool is especially important when a user does not know the exact place name but knows approximately the locations of interest or when the name of a place has changed.

Nevertheless, while the librarians judged the system to be creative and potentially useful, they were not eager to implement such a system in their own libraries.

Conclusion

New spatial information retrieval tools are needed to improve the efficiency and effectiveness of geographically referenced searching. The GeoMatch prototype demonstrates that a graphic-based interface can mine the geographical data buried in MARC records and other geospatial sources and visualize the new knowledge discovered in these data. Combined with the text retrieval capability, this knowledge discovery tool provides

users with greater flexibility in locating the information they need. Discovering knowledge in geospatial data is distinct from text information searching because it uses algorithms to convert coordinate information into user-understandable and useful knowledge.

The main contribution of GeoMatch is the quantitative analysis of the relationship in the retrieval process. Not only can it help users to more precisely define their information need and adjust the searching strategy, but it can also be used to rank the results.

The study of the MARC format shows that it supports the data requirements of GeoMatch, and no additional information is required for converting an existing online catalog to GeoMatch.

Future research in geospatial information retrieval systems will focus on the usability of the system and the theoretical framework of spatial information retrieval, including:

1. usability testing of GeoMatch to study the user friendliness and usefulness of the system;
2. field testing of implementing GeoMatch in a library catalog system;
3. evaluation of the efficiency and effectiveness of the quantitative overlapping function;
4. design of the formula and algorithms to rank the searching result using factors from spatial comparison and factors from text information retrieval such as keywords;
6. application of such a system to information sources other than paper maps, including electronic images and information that can be geographically referenced; and
7. accessibility of such a system over the Web.

Results from these studies could enrich the theories in spatial information retrieval and lead to more powerful and user-friendly information retrieval tools.

REFERENCES

Bell, D. A., & Guan, J. W. (1998). Computational methods for rough classification and discovery. *Journal of the American Society for Information Science, 49*(5), 403-414.

Burrough, P. A. (1990). *Principles of geographical information systems for land resources assessment.* Oxford: Clarendon Press.

Cheeseman, P., & Stutz, J. (1996). Bayesian classification (autoclass): Theory and results. In U. M. Fayyad (Ed.), *Advances in knowledge discovery and data mining* (pp. 153-180). Menlo Park, CA: AAAI Press.

Cobb, M. A., & Petry, F. E. (1998). Modeling spatial relationships within a fuzzy framework. *Journal of the American Society for Information Science, 49*(3), 253-266.

Environmental System Research Institute. (1991). *Understanding GIS.* Redland, CA: ESRI.

Fayyad, U. M.; Piatetsky-Shapiro, G.; & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad (Ed.), *Advances in knowledge discovery and data mining* (pp. 1-34). Menlo Park, CA: AAAI Press.

Glossary. (1995). Retrieved August 18, 1999 from the World Wide Web: http://www.libraries.rutgers.edu/rulib/abtlib/alexlib/glossary.html.

Gluck, M. (1995). Understanding performance in information systems: Blending relevance and competence. *Journal of the American Society for Information Science, 46*(6), 446-460.

Larson, R. R.; McDonough, J.; O'Leary, P.; Kuntz, L.; & Moon, R. (1996). Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science, 47*(7), 555-567.

Mangan, E. U. (1984). *MARC conversion manual—maps: Content designation conventions and procedures for AACR2.* Washington, DC: Library of Congress.

Schmitz, J. (1990). Coverstory—automated news finding in marketing. *Interfaces, 20*(6), 29-38.

School of Information Studies, FSU. (1999). *Foundations of information studies.* Retrieved May 17, 1999 from the World Wide Web: http://slis-one.lis.fsu.edu/courses/5230/.

Smith, T. R. (1996). *A brief update on the Alexandria digital library project—constructing a digital library for geographically-referenced materials.* Retrieved August 6, 1999 from the World Wide Web: http://alexandria.sdc.ucsb.edu.

Smith, T. R. (1998). *Alexandria atlas subteam.* Retrieved August 6, 1999 from the World Wide Web: http://alexandria.sdc.ucsb.edu.

Trybula, W. J. (1997). Data mining and knowledge discovery. In M. E. Williams (Ed.), *Annual review of information science and technology* (pp. 197-229). Medford, NJ: Information Today.

Tuzhilin, A. (1997). Editor's introduction to the special issue on knowledge discovery and its applications to business decision-making. *Decision Support Systems, 21*(1), 1-2.

Xu, X. W.; Ester, M.; Kriegel, H. P.; & Sander, J. (1997). Clustering and knowledge discovery in spatial databases. *Vistas in Astronomy, 41*(3), 397-403.

Additional References

Carter, C. L., & Hamilton, J. (1998). Efficient attribute-oriented generalization for knowledge discovery from large databases. *IEEE transactions on knowledge and data engineering, 10*(2), 193-208.

Chen, Z., & Zhu, Q. (1998). Query construction for user-guided knowledge discovery in databases. *Journal of Information Sciences, 109*(1-4), 49-64.

Connaway, L. S.; Kochtanek, T. R.; & Adams, D. (1994). MARC bibliographic records: Considerations and conversion procedures for microcomputer database programs. *Microcomputers for Information Management, 11*(2), 69-88.

Deogun, J. S.; Choubey, S. K.; Raghavan, V. V.; & Sever, H. (1998). Feature selection and effective classifiers. *Journal of the American Society for Information Science, 49*(5), 423-434.

Maddouri, M.; Elloumi, S.; & Jaoua, A. (1998). An incremental learning system for imprecise and uncertain knowledge discovery. *Journal of Information Science, 109*(1-4), 149-164.

Morik, K., & Brockhausen, P. (1997). A multistrategy approach to relational knowledge discovery in databases. *Machine Learning, 27*(3), 287-312.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Librarians and Information Technology: Which is the Tail and Which is the Dog?

HERBERT S. WHITE

ABSTRACT

THIS ARTICLE WILL ARGUE, PERHAPS IN CONTRADICTION to the discussions which precede it, that providing end users with more information does not really address their problems and, in fact, does not even identify them. Users want information in order to do other things, and this means that they must not only have the best information, but also not have it buried in quantities of other information which may be wrong but are more likely to be irrelevant and thereby misleading. Most importantly, our users need some assurance that what they found is the best that could be found. Dealing with these concerns does not require access to more information, it requires a process to sift the chaff from the wheat. Computer programs used by the end user cannot do this, but computer use by qualified information intermediaries on behalf of, and to protect, the end user can. This growth of specialists has been consistent for any field in which both complexity and options have increased, and the suggestion that computers can be programmed to do their own self-filtering effectively is at best naïve. Peter Drucker has predicted that the most important profession in the next century will be knowledge workers, and knowledge workers are not the same as computer systems specialists. The most competent ones are likely to be reference librarians using sophisticated hardware and software, tools which the end user does not know how to use.

This entire issue of *Library Trends* deals with knowledge discovery or data mining, a relatively sophisticated application of electronic databases. However, most database use is not sophisticated, particularly through CD-

Herbert S. White, 330 E. El Viento, Green Valley, AZ 85614
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 264-277

ROM and the Internet. This puts databases increasingly into the hands of people who are ill-equipped to search them, but who do not necessarily know how ill-equipped they are. Unfortunately, the impression has been created that *anyone* can find not only the right information but also the "best" information by simply sitting down at a computer terminal. Librarians have unfortunately promoted and encouraged these misconceptions by their own insistence that end users search for themselves and to stop bothering the "busy" librarians. In this exercise, end users may or may not find the "correct" information, but they may also find huge quantities of information which are, for them, irrelevant or misleading. End users will then use whatever they found without ever knowing because we refuse to use our expertise to help them. Based on my own experience over the past half century in dealing with a wide range of information problems and services, I will use this article to point out the problems inherent in such simplistic and abdicative approaches.

I made the decision to become a librarian during my junior undergraduate year as a chemistry major in 1948. Part of the reason was my growing awareness that I probably faced very little of a future as a chemist except by working in a laboratory, and I didn't really want to do that for the rest of my career. The other reason came from the growing realization that neither chemistry students nor chemistry professors really knew how to find information in a university library. They would find "something" and make do with that. Whether they had found the best information or all of the correct information they would never know, although they would never admit that they had not found everything they should have found. Students were occasionally caught in that deception, faculty never were. All research reported from the literature was claimed to be complete, and that claim was simply accepted as true. At the time, I knew virtually nothing about librarianship, except for the observation that most librarians were humanists and had not the vaguest idea what chemists were talking about, but that they discouraged such conversations in any case. Researchers "were supposed to" find their own information. If faculty, but particularly students, were helped in anything but the most simplistic directional assistance, we were simply encouraging sloth. While I did not really know what librarians did, because I don't ever recall using my high school library, I was blessedly unencumbered by that ignorance. I only knew what I felt librarians should do, at least for chemists, although I learned quickly that it also applies to other fields. Librarians could and should find the correct information to meet the specific needs of each patron, in part because these individuals were untrained and incapable of finding it for themselves, but primarily because they would rarely if ever admit that shortcoming. Students sometimes get caught in providing incomplete and erroneous information, particularly if the instructor only assigns what he or she already knows. Working professionals are rarely

caught in that deception, and the higher their level of prestige and importance, the safer they became. Indeed, if what they claimed to have found from their "research" was totally unintelligible to others, their claim to brilliance was safest of all.

I had no way of knowing then how correct my totally unsupported hypotheses were but, in the almost half century since becoming a librarian, almost equally divided between operational management and administration in the area of scientific information and the academic pursuits of academic research, teaching, and administration, I have learned the truth of my assumptions many times over. What has surprised me, and continues to surprise me, is the passionate unwillingness of many, if not most, librarians to assist the foundering (even if unconfessed) client to find what is really needed to meet an information need. Thus librarians who could carve out, particularly in an age of computerization, that which geometrically magnifies the amount of information (both useful and useless for the individual need), the crucial role of what I call information lifeguards and Peter Drucker calls knowledge workers, stubbornly refuse to do so. They prefer to handle administrative and clerical details, to build "gateways" to knowledge, and in any case never to intrude into the researcher's right to founder, thereby leaving us with the "rights," as one researcher invited to speak at a sponsored library conference suggested, to build boutiques of information and, when necessary, sweep the floors (Rockwell, 1997). This is certainly not any sort of professional agenda which a real profession, as described by Andrew Abbott (1988), would select for itself. There will be more about Abbott's premise and our failure to seek a road other than an insistently clerical one later in this article. For the moment, it will suffice to note only that this strange reluctance to take professional responsibility for what we presumably know perhaps uniquely, but certainly better than our clients, serves neither them nor us. It is a philosophy wrapped in the professionally self-deprecating "give 'em what they want," and I have suggested to medical librarians that our practice of simply showing clients to terminals and explaining how to use them without any attempt to determine how well they did in meeting their own needs amounted in their field to an encouragement of self-medication. It was the equivalent of saying to patients "Here is the pharmacy. Help yourself to whatever you want."

I learned very quickly, as a sci-tech librarian at the Library of Congress, the Atomic Energy Commission, and the aerospace industry, and well before my introduction to information technology by coming to work for IBM and later NASA nine years into my professional experience, that the scientific literature grew even more rapidly than I had assumed. Statements that each day enough scientific articles were written to fill several complete sets of the Encyclopaedia Brittanica may have been inaccurate or even apocryphal but, even if they were close, they confirmed the

impossibility that any individual, but particularly individuals who sought necessary information for the purpose, not of its own virtue but to be able to do something else with it, a process which would also require time, was doomed to fail. On a more specific note, I recall that, during my own vice presidency at the Institute for Scientific Information in the early 1970s, this company annually announced and described 200,000 new organic compounds, and that wasn't even all of them, only the most important ones. The literature growth in other fields may not have been as dramatic but, in any case, the "ease" of accessing information on computer terminals with which all now live has re-magnified this problem. Technology, whether in databases, listservs, or e-mail, brags about the large quantities of information we now receive. Whether or not it is good information is our problem as end users and, of course a growing problem as technology becomes "more efficient" in quantifying our access. Could librarians help here? Has it occurred to them to offer? Drucker has noted that, in his view, the most important profession after the start of the new millennium will be knowledge workers. Who are these people going to be? Drucker does not specify, but might they perhaps be the information lifeguards I call reference librarians? Or do they all have to have MIS degrees?

I am indebted to my long term colleague Herbert Brinberg (1986) for a cogent and simple definition of why different groups of people need and want information, at least in a professional setting. Chat rooms, playing solitaire online, and browsing for pedophilic and pornographic literature does not count, at least within the context of this article. Brinberg argued that basic pure researchers wanted only raw materials which they would then sift for themselves. Applied researchers and operational workers wanted specific answers to detailed questions. Upper level managers needed to know what their decision options were, and the implications of these options. Brinberg noted, quite correctly, that these different users required approaches suited to the individual need and not some overall policy. Some clients want only minimal help, others would happily turn the entire problem over to a librarian, if it is a librarian they trust. Twenty years in corporate information work has taught me that.

Librarians tend to treat all clients as though they were basic researchers, who only want to be pointed at information sources, although this is particularly true in academic libraries. However, even in the most prestigious institutions, there is very little basic research going on. This has been noted by such diverse sources as the *Chronicle of Higher Education* and humanist scholar/librarian Charles Osburn. My own confirmation comes from the Institute for Scientific Information's publication of "Who is Publishing in Science" (WIPIS). During the years (1970-1974) when I was connected with this publication, fully half of the authors cited for publishing in the literature wrote only one publication and never again. Even when they wrote more than one, it might well be the well-known process

of extending one particular piece of research (such as a dissertation) into as many satellite articles as possible.

However, even if we prefer to deny the premise, well supported as it is, that only a few faculty members do a great deal of research and publishing, a great many others, particularly after they achieve tenure, do very little or none at all. However, even this research tends to become applied research, in the social sciences and humanities as well as the physical sciences, particularly because of the increasing influence of government grants and contracts. Such work is applied precisely because it seeks to "prove" what the funding application postulated. Disproving your own hypothesis might be honest, but it would endanger the chances for additional funding. Most research is then decidedly applied because it seeks to accomplish two things: (1) validate the hopes expressed in the funding application, and (2) demonstrate the need for additional funding. Most "research," including academic research, does not seek raw materials. It seeks "proof" for what we already "know" to be true. The finding of contrary information, whether by the researcher or a librarian, is not always accepted graciously. As noted earlier, we can not only pretend that we found all of the needed information, but also that the conflicting data we did find was not found at all. This is not intended to be cynical, only an accurate observation. In all of my years in the corporate and academic sectors, I know of no scheduled policy or decision making meeting which was ever postponed because the literature review was incomplete. We have what we have by the deadline, and whatever that is we claim to be enough.

If librarians fail to serve applied researchers within the framework in which they work, they tend not to serve the administrators who seek to know what their options are at all, and we must remember that not only in industry but also in universities there are powerful administrators who long ago stopped doing research, if indeed they ever did research, but who in any case make policy decisions which affect the status and operation of libraries and librarians. Why librarians adopt stances and policies which are so consistently counter-productive is outside the scope of expertise of this writer and perhaps belongs instead in the field of psychology.

One thing we have long observed about any information system users is that they want what they want, and they object to having this cluttered by what they do not want. Not all of them, of course, and it is observation that suggests that no library reference service policy is ever totally appropriate. Different people want to be served in different ways, and the good thing is that, if we ask them how they want to be served, they will tell us, although that only works if we don't edict policies which label those who really want to be helped as either selfish or lazy. If that occurs, they will perhaps do the work themselves or more likely abdicate it to an assistant or secretary, or most likely pretend they didn't really need to know. That option is still open to them, as indeed it was in 1948.

However, one thing we should understand, because it is confirmed by operations research studies, is that individuals find the ideal information file to be the one that contains everything they want and nothing else. Faculty members who remove library books they might want to use again to that most relevant of all small files, their own offices, understand this instinctively. Since it is not usually possible to create an ideal world in which we have everything we want and nothing else, individuals react differently to the dilemma. In the 1960s, when I managed one of the earliest selective dissemination of information (SDI) systems for 600 NASA scientists, engineers, and contractors, we found that some individuals happily tolerated lots of "garbage" to make sure they received everything they really needed. Others, who already felt they received too much, bridled at even one notification which they considered as outside their area of interest. We fine tuned profiles to meet these ranges of individual preferences. That phenomenon of individual difference in preference exists today, even as librarians, and to some extent information technologists, insist that one size fits all as we buy information off the rack.

If individuals who work for a living and need information in order to do something with it have not changed, then of course what has changed has been the growth of a technology which brings more information directly to people more easily and more rapidly. It can even be argued that the provision is also more economical. What is not more economical is the human process of sifting out the chaff from the wheat, no matter how many clever software programs are developed. If this sifting is to take place, who should do it? The more greatly stressed, untrained, and probably more highly paid end user? Or one of Drucker's specially prepared, and often more lowly paid (at least in the case of librarians), knowledge workers or information intermediaries?

We can see an increasing reliance on intermediary specialists in many fields, if not in this one. Many of us recall the days when individuals spent Saturday afternoons working on their cars, including carburetor adjustments. Improvements in automotive technology, obviously for our own benefit, now make this impossible, although it is argued that the inconvenience of having to take our cars in for diagnosis and service is far outweighed by the advantage of having better performing cars. We have also seen this increase in specialization in fields such as medicine and dentistry. My regular dentist recently sent me to an endodontist for needed root canal work because he did not specialize in endodontistry. That individual, finding he was unable to save the tooth, sent me to yet another specialist for the extraction. We can certainly recall when one dentist would have done all the needed procedures.

The examples of automotive mechanics and dentistry are only two of what is really a wide range of examples which could be cited to demonstrate the growth of service specialist professions throughout the economy

to allow us to take advantage of the greater opportunities and options which more complex technology, in all areas, now affords us. As opportunities become greater and procedures more complex, we rely increasingly on specialists, and economists confirm that the service sector—the people who do for us what we are now either incapable of doing or unwilling to do is the most rapidly growing field not only in the United States but in the developed world. That the particularly emphatic changes, growth, and complexity in the information sector should have given rise to a swelling cadre of what Drucker calls knowledge workers, and what I prefer to call information intermediaries or simply reference librarians, seems completely obvious. Indeed, it was obvious to Drucker, and his prediction may yet turn out to be completely true. The growth of management information systems (MIS) as an academic discipline is just one example of this phenomenon. However, what is disturbing, at least to me, is that the emphasis here is not on adapting machines to people, it is rather adapting people to machines. The extent to which this has now become the operational mantra of what once were called our library education programs may simply confirm not that the philistines are at the gate, but that they are in charge of our institutions. Certainly the emergence of a new class of educators in our fields, who not only have no idea of what libraries are and do, but who also see no need to learn, tends to confirm this fear.

Why has the development of highly paid specialists who help the general public deal with new options, opportunities, and complexities, completely bypassed this field? How is it possible that, as both the quantity and the importance of information grow at a rapid rate, the number of reference librarians in academia, government, and industry declines (Abbott, 1988). It occurs to me that there are at least three reasons. The first is the fact, first noted by me in 1948 and since repeatedly confirmed, that information ignorance does not need to be admitted and is usually not admitted. Whatever we have is "enough." How, indeed, could we admit that we don't know anything? As a consultant in the assessment of corporate libraries and information centers, I have found quite a few which were inadequate, some whose librarians realized they were inadequate, but none whose users felt their library service was bad. What complaints they utter concern collection access, but even these criticisms are muted. The reason is obvious. If I am doing a good job, and deserving of promotion, salary increases, and grant funding, I must first state confidently that I am doing well and that somehow I know everything I need to know. The process is not as simple in automotive repair and dentistry, because a car which still does not run, or a tooth which still hurts denies the premise that everything is fine. Since end users and upper management either genuinely do not know or at least refuse to acknowledge how inadequately information processes serve them, it is incumbent that the people who presumably do know, the professional

librarians, make the point not of how wonderful libraries are but rather of how inadequate they are and how good they could be. That librarians suicidally never made this point is, however, a part of my third reason and will have to wait.

The second reason comes from the incessant propaganda with which the developers and sellers of computer systems, both hardware and software, constantly bombard us. These messages tend to fall into three categories: (1) using technology is easy, (2) using technology is fun, and (3) using technology saves both time and money. This article will only cite one example of each of the first two because, thus prompted, the reader can certainly find his or her own. The best example for me of the argument that the use of technology is easy comes from a frequently aired television commercial for America Online. In it a young man urges his friend to come with him to a basketball game. The friend declines. He cannot go because he has to order airplane tickets, he must send a birthday present to his mother, and because his child needs to go to the library to locate information on dinosaurs. The friend reassures him that this is all "easy" with America Online and, as we watch in admiration and fascination, the tickets are ordered, flowers are dispatched to his mother, and the printer disgorges pages of encyclopedia information about dinosaurs. What the child is supposed to learn from all of this is not clear, but it is assumed the viewer will not notice.

The second example of the point that using computers is "fun" is best demonstrated for me in a commercial for Hewlett Packard, which demonstrates ingenuity which I consider very effective. We are ushered into the plush office of a very busy executive through the use of an unobtrusive camera. We know he is an executive because the office is so large and tastefully furnished; we know he is busy because it is late at night and he is still hard at work on his Hewlett Packard computer. He rejects our interruption by stating that he is very busy and has no time. Suddenly he moans in anguish. When asked solicitously what has gone wrong, he replies that he has hooked his tee shot into the lake. We all know, and managers have learned, that computer terminals behind a closed door are a potential for doing work and also for wasting time and playing games. Hewlett Packard would be foolish not to stress this second feature, because it probably sells at least as many computers. I am not criticizing either company here for doing what obviously is intended to sell computers. That is their primary responsibility to their stockholders. The problem is not only that these advertisers have a lot of money (I haven't even mentioned Microsoft), but primarily that there is no counter-strategy by those most negatively affected, librarians.

The third argument, that the use of computers saves money, is of course nonsense, and even IBM had stopped stressing this advantage way back when I worked there in the early 1960s. It is both foolish and

unnecessary to claim that computer technology is cheaper, when it is far easier and far more important to demonstrate that the proper use of technology (and even some of the improper use) is cost effective. However, it is wrong to make straight cost comparisons, because this would be a comparison between apples and oranges. Indeed, the use of technology is clearly potentially cost effective in libraries, primarily because it allows that far more effective work be done. However, we must deal not only with the additional hardware and software costs, we must also deal with additional professional staffing costs to use the advanced technological opportunities more effectively. That is why I now have the "privilege" of paying three dentists instead of one. It is good for my dental health. I am certain that corporate, government, and academic administrators really understand this as well, but perhaps I am wrong. Certainly librarians have made no attempt to make a point which should be easy to make—access to more information by more highly paid people who don't really know what they are doing costs more. Obviously.

The mystique that somehow having computers is enough to assure success in information and in education is perhaps best exemplified by the present federal argument, expressed by Vice President Albert Gore, that the solution to our educational problems is making sure that all school children have computers. Presumably not librarians, because they are not necessary. Learning to use computers is both "easy" and "fun."

All of which brings me finally to the third reason. I have always understood, in many years working with information technology, that vendors prefer end user searching to librarian searching. End users have more money, there are more of them, and because they search more sloppily they will spend more. I do not resent this strategy because it makes sense—for them. However, silent acquiescence makes no sense for us.

The great problem for this profession is the lack of any sort of professional philosophy about what libraries are and what librarians do. The issues are no longer discussed in our professional literature, and our library education programs have moved away from any consideration of institutional management. Instead, we have become survivors trying to cope under a barrage of budget cuts which never consider the implications of those budget cuts simply because nobody makes upper management face them. As an adjunct professor at the University of Arizona School of Information Resources and Library Science, I now teach a course in planning and evaluation precisely because I am painfully aware that, to an overwhelming extent, librarians do not plan. Instead, they react to what others have already decided about the future of the library. Planning, by contrast, is an early process of pointing out to upper management the alternative implications of various decision options before those decisions are made. Librarians have largely abdicated any confidence that they understand what they ought to be doing far better than anyone

else, the essence of any professional discipline. Thus, the Baltimore County Public Library motto of "give 'em what they want" rather pathetically sums up the vision of many librarians. It is not "give 'em what they need" or even "make 'em aware of what they could have and should have."

Nor do libraries really evaluate. Instead, they rig questionnaires which only ask people already in the library, and therefore an obviously biased constituency, how they "like" their library. As compared to what? The responses may be predictable, but they are also not only useless but dangerous when we recall Drucker's injunction that the essence of management communication is exception reporting—what ought to be happening but is not happening.

There are three distinct roles that libraries can play, and the later named ones are far more important and offer far more potential than the earlier ones. The first is the library's role in recreation. It is the easiest to explain and to justify, and it is indeed the role, particularly for public libraries, that our clients most easily identify. It is also, of course, the most trivial and becomes the most dangerous during the budget review procedures which have become standard in all management operations. These reviews force the ranking of priorities, and recreational activities (parks, libraries) can never compete against the priorities of police protection, road repair, and public health. When libraries are judged in this environment, the evaluation usually comes out as "of course we favor good libraries, but. . .". In the context of the information world, this sometimes comes out as "of course information is important, but what has this to do with libraries?"

The second role, in education, is the one which probably the majority of librarians embrace. In this context, we don't so much answer questions as teach students to answer their own questions. It is certainly a different approach from that practiced by plumbers and mechanics who are not likely to teach us how to fix our own leaks and transmissions. However admirable one might consider this role as an objective, it cannot succeed as long as the "other" educators, be they teachers or professors, fail to acknowledge us as partners of equal importance.

This trivialization of our educational role can be easily seen, on the one hand, in the willingness of teacher unions to sacrifice librarians to retain teaching slots. On the other hand, we must recognize the failure to grant (as at institutions like Harvard) faculty status to librarians, and the constant pressure to take both faculty status and tenure away from librarians. That pressure sometimes comes from administrators, but I have failed to see it ferociously opposed by the American Association of University Professors (AAUP). Finally, the failure of our "fellow" educators to accept us as full brothers and sisters can be observed in the traditional low, almost invisible, status of library programs and particularly library research within the federal Department of Education and the research-oriented

Institute for Education. Educators have now neatly finessed this problem by transferring library programs to the Institute of Museums and Libraries, again with us in the junior positional listing and under the directorship of museum experts. And yet nobody in this profession, in its leadership, and in its professional publications finds this objectionable, let alone intolerable.

The third role, that of information intermediaries, is clearly the one which, in this age of growing information output, growing information access, and therefore growing information confusion, poses the greatest potential for this field, as Drucker recognized in his stressing of the importance of knowledge workers. However, acceptance of role number three causes a potential direct conflict with role number two, that of educators. In role number three we do not teach end users to solve complex problems without us, even as that educational exercise is at best problematical because we do not know whether such users whom we have turned loose in the information ocean ever find what they need—no I did not say want. WANT is, particularly for the unprepared, as irrelevant here as it is in medicine. In accepting our roles as information intermediaries, we seek rather to make our clients dependent on our unique expertise. To place this into the context of a profession's responsibility and sense of expertise, I will now return to the writings of Andrew Abbott (1988), briefly mentioned at the beginning of this article, and his definition of a profession. Professions, Abbott argues, have the unique responsibility of addressing human problems amenable to expert service, and I interject only to note the words *problems* and *expert.* Abbott continues that professionals compete vigorously for existing and newly emerging problem jurisdictions, and that they strive to expand those jurisdictions by preempting the activities of other professions.

The reader can certainly understand what sort of expanding jurisdiction, as well as amenable problem areas, computerized access to information represents, and it should be equally obvious what the other fields are at whose expense we should be expanding our jurisdictions. That the growth of computer-based information access not only provides opportunities but also changes the ground rules is certainly clear today. Indeed, it has been clear for thirty-five years.

In 1964, in what can be argued to have been the very beginning of the technological information age, my friend and mentor, Mortimer Taube (1964), the president of Documentation Incorporated, noted that the development of the MARC system by the Library of Congress, and its reliance on what is now seen as rudimentary but was still exciting technology, allowed librarians to rethink and completely restructure their cataloging rules, particularly with regard to subject analysis. That analysis, Taube noted, was constrained by the economic problems of having to file 3" x 5" catalog cards, and this limited subject analysis to the perfunctory level of

perhaps one or two broad subject headings. Even the computer technology available in 1964 removed that limitation and allowed for analysis in far greater detail. Taube expressed the concern that the library profession would fail to see this opportunity and simply devise techniques for computerizing the Anglo-American Cataloguing Rules. And that, of course, is exactly what we did do.

However, by far the greatest opportunities thirty-five years later lie in the expanded role for reference librarians to claim for themselves Abbott's territorial role in doing what others should not do and, more importantly, could not do. That we have failed to seize this opportunity is most evident in the decline in the number of reference librarians, even as we are deluged by reports of growing information files, growing information needs, and growing information complexity. Justifying additional reference librarians as the most cost-effective strategy for dealing with this issue should be relatively simple. However, we continue to see the strategy of National Library of Medicine administrators of teaching medical practitioners to search for their own information online, even as we are also told that the development of Health Maintenance Organizations (HMOs) increasingly turns the doctor into an overworked production employee with neither time nor energy for undertaking information searches at the end of a fourteen hour working day. Since medical librarians are both much cheaper and better trained for information searches, the solution should be obvious, yet no one sees and no one clamors for it.

In the absence of management courses in our library education programs, in the lack of professional discussion concerning our management strategies, and in the absence of research literature on this topic, it is difficult to understand why librarians insist on following a suicidal policy of shifting professional duties from their own desks to terminals to those of the end user, while they retain the routine activities which make them look like clerks. And yet they do. In reviewing grant funding proposals for the Institute of Museums and Libraries, I found numerous requests for additional money with which to purchase hardware and software for our end users. There were no proposals for funds to purchase tools to be used exclusively by librarians, to give them skills end users could never possess, and to make them more important. These are not disciples of Andrew Abbott.

Just as Peter Drucker predicts, the growth in the role of information intermediaries or knowledge workers is certain, even as the part which librarians will play is not nearly as certain. Once we get past our fascination with teaching children to play computer games on the premise that playing on computers is somehow more virtuous than playing soccer or basketball, and once we understand that having adults waste time on computers playing solitaire, surfing the Net aimlessly, or downloading

anything for any reason or for no reason is more educational than watching soap operas on television, we will be left with the information needs of people who work for a living, and who need information in order to do this work. Herbert Brinberg (1986) has given us a clear indication of who these people are.

In addressing information needs of end users, there are two things we need to keep in mind. The first is that here, as in any other segment of society, we delegate what we can delegate, and save for ourselves only what we must do ourselves. The development of terminals in executive offices has not reduced the number of administrative assistants, precisely because having more assistants at our beck and call makes us more powerful. The second is that ignorance does not need to be admitted. Complete knowledge will be claimed whenever an admission to the contrary gets in the way of the primary objective.

For end users to delegate to information intermediaries, there are still two additional requirements. The first is that the user must trust the intermediary. Trust cannot be simply claimed, it must be earned. However, once it is earned, it is freely and openly given. Good reference librarians, whom clients insist on using even if they have to wait until they come on duty, understand this and appreciate this, and their bosses should also understand that clients usually know who the good librarians are. The second requirement is one of convenience. Clients want to be helped on their schedule and not the institution's. However, technology can be very helpful here. American Express learned long ago the virtue of establishing an 800 number telephone staff twenty-four hours per day. Whoever answers the phone has complete access to your file and can help you. The Social Security Administration has learned the same thing. Its 800 number is staffed from 8 A.M. on the East Coast until 6 P.M. in Hawaii. You never get the same person twice, but it doesn't matter. The person who answers the phone is well trained, has complete access to your file and organizational policies, and can either put you on hold or call you back while he or she seeks either clarification or approval from a higher level of management. Is this possible for an online reference service? Of course it is!

Given acceptable options, clients will treat the increasing opportunities and options in information access exactly the same way they treat increased complexity in automotive repair and financial investment decisions. We delegate to a specialist whom we trust, and who will work within our time frame. A high level executive made the point quite clearly. He was delighted at the improvements in air transportation, which now allowed him to fly far more rapidly without the delay of refueling in a luxuriously appointed corporate jet. However, that did not prompt him to learn how to fly—not as long as he could hire a qualified pilot.

References

Abbott, A. (1988). *The system of professions.* Chicago, IL: University of Chicago Press.

Brinberg, H. (1986). *Unpublished talk presented at the Conference of the International Federation for Documentation (FID).* Copenhagen, Denmark. September.

Rockwell, R. C. (1997). Using electronic social science data in the age of the Internet. In L. Dowler (Ed.), *Gateway to knowledge: The role of academic libraries in teaching, learning, and research* (pp. 59-80). Cambridge, MA: The MIT Press.

Taube, M. (1964). *Unpublished talk presented at a meeting of the Washington, D.C. Chapter of the Special Libraries Association.* May.

About the Contributors

HELENA AHONEN is a faculty member in the Department of Computer Science at the University of Helsinki, Finland. During the academic years 1997/98 and 1998/99 she worked as a postdoctoral researcher at the Wilhelm-Schickard-Institut of the University of Tübingen, Germany. She is the author of several articles on document management and knowledge discovery in databases.

GOBINDA G. CHOWDHURY is a Senior Lecturer in the Division of Information Studies, School of Applied Science, at Nanyang Technological University, Singapore, where he teaches and guides research in different areas of information retrieval, knowledge organization, information sources and searching, and digital libraries. He has been teaching library and information science for the past thirteen years and has taught at various universities in India, England, Africa, and Singapore. He has wide international teaching and research experience and over fifty publications that include four books and a number of research papers that appeared in a number of reputed journals, seminars, and conferences.

KENNETH A. CORY is currently an instructor and media specialist at St. Thomas More Academy in Burton, Michigan.

BIPIN C. DESAI is a faculty member in the Department of Computer Science, Concordia University, Montreal (Canada). His research interests are in the areas of database systems, application of AI to information systems, and the virtual library. He is responsible for the CINDI virtual library project. He has published numerous articles, is the General Chair of IDEAS (International Database Engineering and Applications Symposium), and has written a popular textbook on database systems.

QIN HE is a doctoral student in the Graduate School of Library and Infor-

mation Science at the University of Illinois at Urbana-Champaign. Her particular interests include knowledge discovery and data mining based on semantic analysis.

BARBARA H. KWASNICK is a faculty member in the School of Information Studies at Syracuse University, where she instructs in the areas of knowledge representation and organization and research methods. She is active in the American Society for Information Science Special Interest Group on Classification Research (SIG/CR) and was the co-editor of the first five volumes of *Advances in Classification Research*, the proceedings of that SIG's annual workshop. Ms. Kwasnik is also a member of the International Society for Knowledge Organization where she has served on the program committee.

F. W. LANCASTER, editor of Library Trends and Professor Emeritus of Library and Information Science at the University of Illinois at Urbana-Champaign, has been working in or around libraries for almost fifty years. He is author or co-author of eleven books (several of which have earned prestigious national awards) and editor or co-editor of twelve others. He has lectured at more than seventy universities or colleges in sixteen countries.

M. JAY NORTON is an Assistant Professor in the School of Library and Information Science at The University of Southern Mississippi where she teaches courses dealing with database construction and applications, information science, computer applications in libraries, library automation, and media utilization. As a member and chair of the American Society for Information Sciences Computer Retrieval Services special interest group, she has been involved in researching retrieval systems from a wide variety of perspectives. She has recently completed a book, *Introductory Concepts in Information Science* to be published by Information Today, Inc. Ms. Norton also serves as a research and information technology consultant.

MARIA PINTO is Professor in the Documentation Faculty of the Granada University, where she teaches courses on information processing and management of quality in library and information science. She is the author of six books (two in second editions) in areas related to knowledge representation, content analysis, abstracting methods and products, and the role of quality management in information processes. Ms. Pinto has also published chapters in monographs and articles in international reviews, one of which received the MIP Award of the FID as the Best Article of the 1994 year. She has participated as a partner in Project I+D financed by the European Community and has been responsible for investigation projects financed by the Education Ministry of Spain.

JIAN QIN is Assistant Professor at the School of Information Studies at Syracuse University in Syracuse, New York. She was the recipient of the OCLC LIS Research Grant in1997 and the ISI Citation Research Grant in 1997. Ms. Qin is the author of over twenty journal articles, technical reports, and conference papers dealing with scientific communication, metadata, keyword semantic pattern analysis, and bibliometrics. More recently she co-edited a topical issue on Web research and information retrieval for *Information Processing and Management.*

NADER R. SHAYAN, a graduate student in the Department of Computer Science, is currently working in Montreal as a consultant.

RAIJAN SHINGHAL is a faculty member in the Department of Computer Science, Concordia University in Montreal, Canada. He is the author of the book *Formal Concepts in Artificial Intelligence* published by Chapman and Hall in London.

NEIL R. SMALHEISER is a Research Assistant Professor in the Department of Psychiatry at the University of Illinois at Chicago. His experimental research concerns the role of extracellular matrix proteins in nervous system development, function, and disease. Mr. Smalheiser has collaborated with Don Swanson on the Arrowsmith project for the past seven years.

HENRY SMALL, after a brief career as a historian at the American Institute of Physics' Center for History and Philosophy of Physics, joined the staff of the Institute for Scientific Information in 1972, where he is currently Director of Contract Research. His 1973 paper in *JASIS* on co-citation in the scientific literature led to numerous papers on citation analysis and the mapping of science. Mr. Small's current research centers on delineating document pathways through science. He has served on the *JASIS* editorial board since 1986 and is a Fellow of the AAAS. In 1987 he received the JASIS Best Paper Award and the Derek de Solla Price Medal from the journal *Scientometrics* and in 1998 the Award of Merit from the American Society for Information Science.

DON L. SWANSON is Professor Emeritus at the University of Chicago and is the author of numerous articles related to information science and in particular to what is now the Arrowsmith project which he originated in 1986 with his work on "undiscovered public knowledge."

HERBERT S. WHITE, the author of ten books and over 300 articles and reports, for twenty-five years managed in the corporate and government sectors in such posts as IBM Program Manager, Executive Director of the

NASA Scientific and Technical Information Facility, and Senior Vice President of the Institute for Scientific Information (ISI). Mr. White followed this with twenty years at the Indiana University School of Library and Information Science including ten years as dean. Retired in 1995 as Distinguished Professor, he now resides in Arizona where he continues to teach, lecture, consult, and write including the regular column "White Papers" for *Library Journal.* He is past president of both the Special Libraries Association and the American Society for Information Science, and a former board member of the International Federation for Documentation (FID), the American Federation of Information Processing Societies (AFIPS), and the Society for Scholarly Publishing. A former member of the ALA Council and columnist in *American Libraries,* Mr. White is the recipient of the Melvil Dewey Medal.

LIXIN YU is an Assistant Professor at the School of Information Studies, Florida State University, where he teaches courses in database management, user interface design, and information system design and development. He worked as a Project Manager at Geosocial Resources, Inc. and has been working on Geographic Information System projects since 1990. He has published articles on GIS including "Geographic Information Systems in Library Reference Services: Development and Challenge" *(Reference Librarian,* February 1998) and "Assessing the Efficiency and Accuracy of Street Address Geocoding Strategies" *(Proceedings of GIS '97,* December 1997).

YOUQUAN ZHOU, a graduate student of Concordia, is currently working at IBM (Washington, DC) as a consultant.

COMING IN 1999

Clinic on Library Applications of Data Processing
1998 proceedings

Successes and Failures of Digital Libraries
Edited by Michael Twidale and Susan Harum

PAST PROCEEDINGS ARE ALSO AVAILABLE:

1997 Proceedings
Visualizing Subject Access for
21st Century Information Resources
Edited by Pauline Atherton Cochrane and Eric H. Johnson
$30.00*

1996 Proceedings
Digital Image Access & Retrieval
Edited by P. Bryan Heidorn and Beth Sandore
$30.00*

1995 Proceedings
Geographic Information Systems and Libraries:
Patrons, Maps, and Spatial Information
Edited by Linda C. Smith and Myke Gluck
$30.00*

Send orders to: GSLIS Publications Office, Room 313, 501 E. Daniel Street, Champaign, IL 61820. Prepayment required; Visa, MasterCard, American Express, Discover and checks (payable to the University of Illinois) accepted. Information regarding other publications can be obtained by writing to the above address or can be accessed at our Web site:
http://edfu.lis.uiuc.edu/puboff

*Price does NOT include shipping. Within the United States, the shipping cost is $3 for the first book, $1 for each additional book in the same order. Outside of the United States, the shipping cost is $5 for the first book, $1.50 for each additional book in the same order. (We ship Fourth Class Library Rate.)

INDEXING AND ABSTRACTING IN THEORY AND PRACTICE

2nd edition

By F. W. Lancaster

SECOND EDITION FEATURES

MULTIMEDIA SOURCES AND THE INTERNET

Award-winning author F.W. Lancaster has revised his widely used text to address growing complexities in the field. Featured in the second edition of ***Indexing and Abstracting in Theory and Practice:***

- New multimedia sources chapter
- New indexing within the Internet chapter
- Updated chapters on text searching, automatic processing methods, and the future of indexing and abstracting
- Nine updated chapters on basic principles and theories
- Modified practical exercises

In addition to use as a text, ***Indexing and Abstracting in Theory and Practice*** holds value for managers of information services and others concerned with indexing, abstracting, and all related issues of content analysis.

The Publications Office
Graduate School of
Library and Information Science
University of Illinois
501 East Daniel
Champaign, IL 61820
(217) 333-1359
(217) 244-7329 fax
puboff@alexia.lis.uiuc.edu

Orders must be prepaid to
The University of Illinois
Major credit cards and checks accepted

ISBN 0-87845-102-1
426 pages
cloth
$47.50 plus shipping